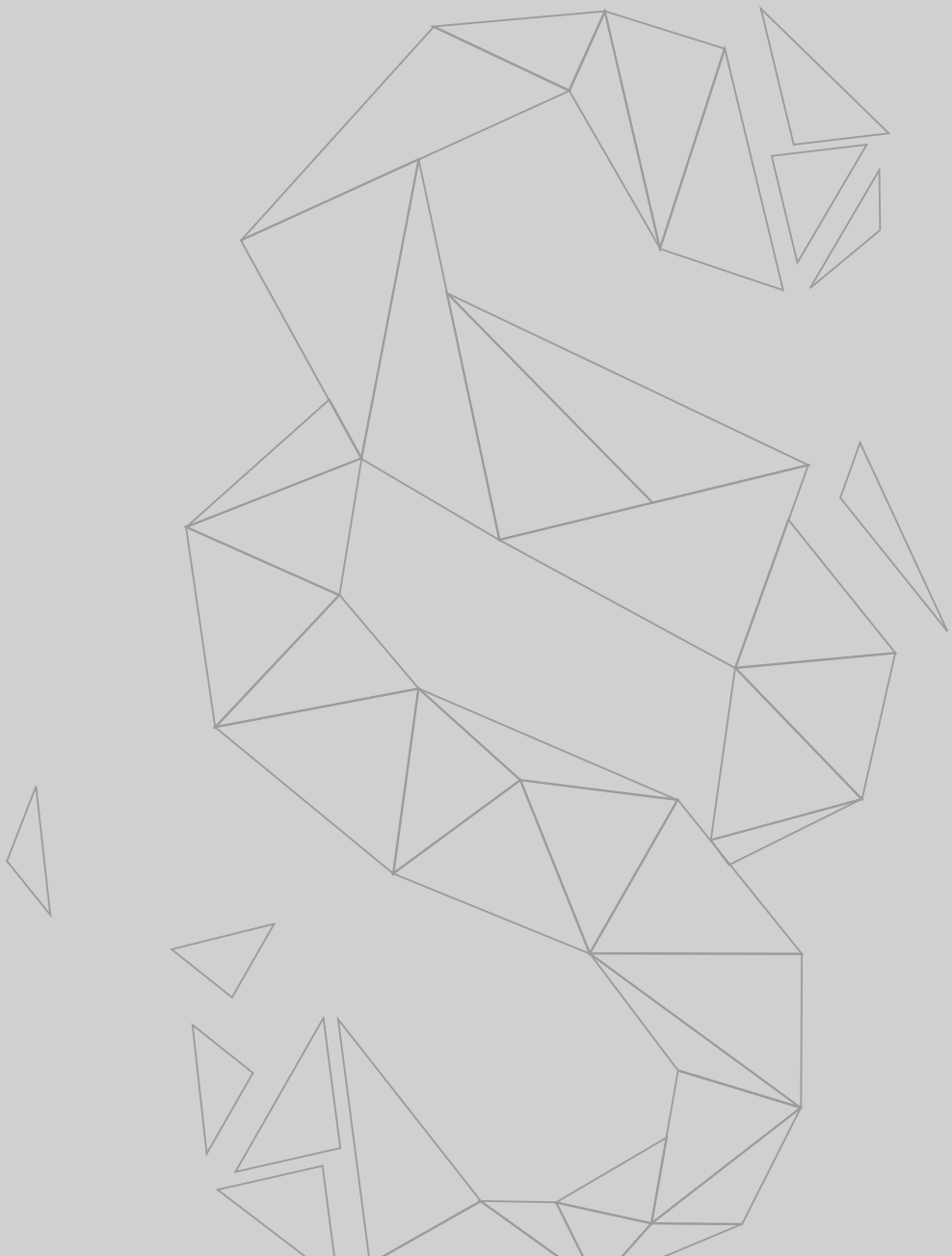
STUDIE

Technische und rechtliche Betrachtungen algorithmischer Entscheidungsverfahren

Gutachten der Fachgruppe Rechtsinformatik der Gesellschaft für Informatik e.V. im Auftrag des Sachverständigenrats für Verbraucherfragen

SVRV Sachverständigenrat für Verbraucherfragen



Zitierhinweis für diese Publikation:
Gesellschaft für Informatik (2018). Technische und rechtliche Betrachtungen algorithmischer Entscheidungsverfahren. *Studien und Gutachten im Auftrag des Sachverständigenrats für Verbraucherfragen.* Berlin: Sachverständigenrat für Verbraucherfragen.

Berlin, Oktober 2018
Studien und Gutachten im Auftrag des Sachverständigenrats für Verbraucherfragen
ISSN: 2365-8436

Herausgeber:
Sachverständigenrat für Verbraucherfragen
beim Bundesministerium der Justiz und für Verbraucherschutz
Mohrenstraße 37
10117 Berlin
Telefon: +49 (0) 30 18 580-0
Fax: +49 (0) 30 18 580-9525

E-Mail: info@svr-verbraucherfragen.de
Internet: www.svr-verbraucherfragen.de

Gestaltung: Atelier Hauer + Dörfler GmbH
Druck: Brandenburgische Universitätsdruckerei

GESELLSCHAFT
FÜR INFORMATIK

Technische und rechtliche Betrachtungen algorithmischer Entscheidungsverfahren

Gutachten der Fachgruppe Rechtsinformatik der Gesellschaft für Informatik e.V.
im Auftrag des Sachverständigenrats für Verbraucherfragen

August 2018

Inhaltsverzeichnis

Abbildungsverzeichnis

Hinweise

Aus Gründen der besseren Lesbarkeit wird auf die gleichzeitige Verwendung männlicher und weiblicher Sprachformen verzichtet. Sämtliche Personenbezeichnungen gelten gleichermaßen für beiderlei Geschlecht.

Executive Summary

Unternehmen treffen täglich unzählige Entscheidungen. Die Optimierung und Automatisierung von Entscheidungsprozessen ist deshalb Kernaufgabe in vielen Unternehmen. Viele Entscheidungen, die bislang durch den Menschen getroffen wurden, können von Algorithmen und algorithmischen Verfahren ebenso, wenn nicht sogar noch präziser, objektiver und schneller getroffen werden. Dies gilt insbesondere für datenintensive und zeitkritische Entscheidungen, die rationalen (formalisierbaren) Kriterien folgen. Hier spricht man von *Algorithmic Decision Making* (ADM).

Dabei können die zugrundeliegenden Entscheidungsstrukturen direkt vom Menschen vorgegeben sein oder von den Algorithmen selbst aus einer bestehenden Datenmenge extrahiert werden. Letztere erfahren durch informatische und informationstechnische Entwicklungen in den Bereichen der Künstlichen Intelligenz und des Maschinellen Lernens enormen Aufwind. Die Verfügbarkeit großer Datenbestände und die zunehmende Leistungsfähigkeit der IT-Systeme machen das Feld zunehmend attraktiver für Unternehmen. So werden ADM-Verfahren unter anderem bei der Bewertung der Kreditwürdigkeit von Personen (Kreditscoring), sowie bei der Preisermittlung im E-Commerce (*dynamic pricing*) eingesetzt. Die datengestützte Beurteilung führt allerdings zu einer standardisierten Beurteilung und einer Informationsasymmetrie zwischen Unternehmern und Verbrauchern. Gepaart mit fehlender Transparenz kann dies zu einem starken Unbehagen in der breiten Öffentlichkeit und in Expertenkreisen führen.

ADM-Systeme können fehlerhafte Entscheidungen treffen, die möglicherweise normative Anforderungen verletzen und damit Rechte Dritter verletzen. Insbesondere sind das Diskriminierungsverbot sowie das Verbot automatisierter Entscheidungen betroffen. So ist beispielsweise nach dem Allgemeinen Gleichbehandlungsgesetz (AGG) eine Benachteiligung, im Sinne einer Diskriminierung, nach bestimmten individuellen Merkmalen nicht zulässig. ADM-Systeme, insbesondere wenn ihnen maschinelle Lernverfahren zugrunde liegen, können jedoch genau solch problematisches Verhalten aufweisen.

Die Koalition aus CDU/CSU und SPD adressiert diese Thematik und will bei Fehlentwicklungen entsprechende regulierende Mechanismen entwickeln. Im Koalitionsvertrag heißt es, dass die Bundesregierung *„zum Schutz der Verbraucherinnen und Verbraucher Algorithmen- und KI-basierte Entscheidungen, Dienstleistungen und Produkte überprüfbar machen [will], insbesondere im Hinblick auf mögliche unzulässige Diskriminierungen, Benachteiligungen und Betrügereien. Wir werden Mechanismen entwickeln, um bei bedenklichen Entwicklungen tätig werden zu können. Dynamische Preisbildung muss Verbraucherinnen und Verbrauchern nach klaren Regeln transparent dargestellt werden."*[1]

Das vorliegende Gutachten wurde im Auftrag des Sachverständigenrates für Verbraucherfragen (SVRV) von einer interdisziplinären Expertengruppe aus Juristen und Informatikern der Gesellschaft für Informatik e.V. zwischen November 2017 und August 2018 erstellt. Es vereint technische und rechtliche Betrachtungen von ADM-Verfahren mit

[1] Koalitionsvertrag zwischen CDU, CSU und SPD zur 19. Legislaturperiode vom 12. März 2018 [https://bit.ly/2oJXYXF].

ihren zugrundeliegenden Systemen und gibt der Politik konkrete Handlungsempfehlungen an die Hand.

Aus **technischer Sicht** stehen bei der Betrachtung zu algorithmischen Entscheidungsverfahren insbesondere Fragen der „Fairness“ und der möglichen Ungleichbehandlungen im Vordergrund, die durch Unausgewogenheit der Daten, direkte oder indirekte Einflussnahme entstehen können. Der Erklärbarkeit von ADM-Systemen kommt entscheidende Bedeutung zu: Obwohl sich Unternehmen und Öffentlichkeit häufig des Bildes eines intransparenten und nicht nachvollziehbaren Entscheidungsvorgangs (*Black-Box*) bedienen, ist dies nicht notwendigerweise richtig. In vielen ADM-Systemen können Entscheidungsstrukturen transparent und nachvollziehbar dargestellt werden. Bei der Analyse des Entscheidungsverhaltens existieren zwei zentrale Methoden, die die Transparenz von ADM signifikant erhöhen: **Testing und Auditing**.

Aus **rechtlicher Sicht** wird insbesondere das Problem der Diskriminierung adressiert. Dabei muss sorgfältig zwischen datenschutzrechtlichen Aspekten und dem Diskriminierungsschutz differenziert werden. Die Datenschutz-Grundverordnung (DSGVO) enthält ein Verbot automatisierter Entscheidungen, das jedoch nur für vollautomatisierte Entscheidungen gilt und umfangreiche Ausnahmen enthält. Diskriminierungen sind im Anwendungsbereich des AGG unzulässig – allerdings ist der Anwendungsbereich des Gesetzes beschränkt. Das allgemeine Deliktsrecht enthält ebenfalls Schutz gegen Diskriminierung – allerdings sind die Voraussetzungen nicht spezifiziert und schwer nachweisbar.

Die Studie zeigt erhebliche Defizite im geltenden Recht auf. Insbesondere ist die Feststellung einer Diskriminierung – die Voraussetzung jeglichen Rechtsschutzes – beim Einsatz von ADM-Systemen in der Praxis nicht gesichert. So ist der Diskriminierungsbegriff noch nicht ausreichend präzisiert und operationalisiert, weshalb eine algorithmische Prüfung, z.B. durch Testverfahren, nicht verlässlich und rechtssicher durchgeführt werden kann. Zwar gibt es vielfältige Forschungsbemühungen im Bereich *Fair Machine Learning*, jedoch existiert noch viel Forschungsbedarf um dort gewonnene Erkenntnisse praktisch zu erschließen.

Durch Tests von ADM-Systemen können Diskriminierungen festgestellt werden, wobei hier der Nutzung quantitative Gleichbehandlungsbegriffe und die Verfügbarkeit qualifizierter Testdaten besondere Bedeutung zukommt. Jedoch enthält das geltende Recht keine Klarheit hinsichtlich der Rechtsfolgen der Durchführung von Tests. Insbesondere fehlt es an Möglichkeiten, die erforderliche Mitwirkung der Betreiber von ADM-Systemen rechtlich zu erzwingen. Die Anforderungen an Testverfahren für ADM-Systeme sind nicht rechtlich determiniert. Damit sind die Verlässlichkeit sowie eine wesentliche Grundlage der rechtlichen Bedeutung von Tests nicht gesichert.

Weder das Datenschutzrecht noch das Antidiskriminierungsrecht kennt eine Ex-Ante-Überprüfung von algorithmischen Entscheidungen. Es besteht also dringender interdisziplinärer Forschungs- und Handlungsbedarf bei der Spezifikation von Maßstäben zur Erkennung von unfairen bzw. diskriminierenden algorithmischen Entscheidungen. Dies würde einerseits die Rechtssicherheit beim Einsatz von ADM-Systemen erhöhen und andererseits klare Anforderungen für die Entwicklung und den Betrieb solcher Systeme schaffen.

Im **internationalen Vergleich** zeigt sich, dass eine bedingte Kontrolle von Fairness in der algorithmischen Entscheidungsfindung mit vergleichsweise wenig Aufwand möglich ist. Diese ist jedoch notwendigerweise immer unvollständig, sektorspezifisch und muss

kontinuierlich angepasst werden. Ein gut untersuchtes Szenario stammt aus den USA: der *Fair Lending Act* zu dem der *Equal Credit Opportunity Act (ECOA)* und der *Fair Housing Act (FHA)* zählen. Diese zielen explizit auf die Vermeidung von Diskriminierung durch algorithmische Entscheidungsfindung im Bereich Profiling und Bonitätsprüfung (Kreditscoring) ab. In den USA werden bereits Test- und Auditierungsverfahren eingesetzt und konkrete Maßstäbe zum Nachweis von Diskriminierung verwendet. Diese sind jedoch immer mit Annahmen verbunden, z.B. statistischen Grenzwerten, die für eine Anwendung in Deutschland expliziter Überprüfung und ggf. einer Anpassung bedürfen.

Nichtsdestotrotz erkennen einige Länder die zunehmende Bedeutung von überprüften ethischen und rechtskonformen ADM-Systemen an. So versucht eine Regierungsinitiative in Großbritannien „ethisch zertifizierte Algorithmen" zu einem internationalen Marktvorteil der IT-Branche zu entwickeln, der neben einem „unique selling point" auch Verbrauchervertrauen und damit Akzeptanz erhöhen soll. Garantierte Rechtskonformität kann und sollte zu einem Qualitätsmerkmal werden, dass es Unternehmen erlaubt sich als Anbieter für nicht-diskriminierende ADM-Systeme auszuweisen.

Die Handlungsempfehlungen des Gutachtens lassen sich in folgende Aspekte gliedern:

- Es besteht grundsätzlich **legislativer Handlungsbedarf** beim Einsatz von ADM-Systemen und algorithmischen Entscheidungen. Jedoch sind der Umfang des Regelungsbedarfs und die Möglichkeiten der Gesetzgebung derzeit noch nicht deutlich absehbar. Dies liegt zum einen daran, dass das Gefahrenpotenzial von ADM und algorithmischen Entscheidungen noch bei weitem nicht umfassend bekannt ist. Zum anderen bedürfen zahlreiche Rechtsfragen der Klärung. Regulierende Maßnahmen beim Einsatz von ADM-Systemen könnten auch in Form einer Selbst- oder Ko-Regulierung erfolgen.

- **Test-, Auditierungs- und Zertifizierungsverfahren** sind wirkungsvolle Werkzeuge, um rechtsverletzende Diskriminierung durch ADM-Verfahren zu adressieren. Ziel solcher Verfahren muss die Steigerung der Transparenz über die Nutzung von ADM-Verfahren sowie deren Wirkungsweisen sein. Dazu müssen **Standards** entwickelt werden, anhand derer diese Tests und die zugehörigen Audits durchgeführt werden können. Test- und Auditverfahren setzen wiederum die Legitimation durch den Gesetzgeber voraus.

- Die Durchführung der **Tests von ADM-Systemen** ist ein wesentliches Element des Schutzes gegen fehlerhafte algorithmische Entscheidungen. Daher sollten sowohl die Grundlagen von Tests und ihrer Durchführung als auch die Bedeutung von Testergebnissen rechtlich abgesichert werden. Zu den rechtlichen Anforderungen im Einzelnen besteht jedoch noch erheblicher Forschungsbedarf, so dass gesetzliche Maßnahmen erst nach umfassendem Erkenntnisgewinn ergriffen werden sollten.

- Sobald der rechtliche Rahmen für geeignete Testverfahren gelegt ist, sollte eine **gesetzliche Pflicht zur Durchführung hinreichender Tests** eingeführt werden. So können ADM-Systeme vor ihrem Einsatz hinreichend auf Fehler, insbesondere Diskriminierung, geprüft werden.

- Transparenz und Information sind wichtige Schutzinstrumente gegen potentielle Gefahren durch algorithmische Entscheidungen. Daher sollte die **Gewährung von Information** auch durch rechtliche Mittel und entsprechende legislative Maßnahmen sichergestellt werden. Die Einführung von Meldepflichten für Hersteller beim

Inverkehrbringen von ADM-Systeme ist zu erwägen, soweit ein Schutzbedarf besteht. Im Einzelnen besteht jedoch noch erheblicher Klärungsbedarf.

- Zur Einhaltung von Transparenz und Informationspflichten sowie zur Implementierung effizienter und effektiver Test- und Auditierungsverfahren wird die Einrichtung einer **staatlichen Stelle für algorithmische Entscheidungen** empfohlen. Diese muss mit ausreichend Expertise, Befugnissen und Ressourcen ausgestattet sein, die es ihr erlaubt, ADM-Systeme zu testen, zu auditieren und zu zertifizieren. Wesentliche Aufgabe einer solchen Agentur beispielsweise nach dem Vorbild des Bundesamts für Sicherheit in der Informationstechnik (BSI) muss zudem die Steigerung der Transparenz durch Beratung und Information von Entscheidungsträgern in Unternehmen, Verwaltung und Politik sowie der gesellschaftlichen Aufklärung sein.

- Der steigenden Bedeutung algorithmischer Entscheidungssysteme muss die **Ausbildung entsprechender Expertinnen und Experten** an den Hochschulen Rechnung tragen. Betroffen sind vor allem Studiengänge der Informatik und der Rechtswissenschaften. Darüber hinaus gilt es, die Kompetenzen im Umgang mit digitalen Technologien und Daten in der Breite der Hochschulausbildung zu verankern.

- Es bedarf weiterer, vor allem **internationaler und interdisziplinärer Forschungsanstrengungen** an den Schnittstellen zur Informatik und den Rechtswissenschaften (aber auch darüber hinaus), um die vielfältigen offenen Fragen zu adressieren. Spezifische Fragestellungen müssen durch wohldefinierte und konkrete Forschungsprojekte und -initiativen, z.B. durch Stipendien, Promotionen oder weitere Gutachten, untersucht und für den Diskurs aufbereitet werden.

Es liegt im Interesse der Gesellschaft und des Gesetzgebers, klare Regelungen zu schaffen, um Rechtssicherheit für Unternehmen und Bürger zu schaffen ohne Innovationen zu verhindern und gleichzeitig Diskriminierung durch ADM-Verfahren zu vermeiden. Es ist nicht absehbar, dass die Herausforderungen bei der Regulierung von algorithmischen Entscheidungsverfahren in naher Zukunft vollumfänglich beantwortet werden können, insbesondere im Hinblick auf die starke Dynamik im Bereich der Technologie rund um die künstliche Intelligenz bzw. das Maschinelle Lernen. Umso wichtiger ist es, dass die Fragen bereits heute adressiert werden, um dieses zunehmend relevanter werdenden Herausforderungen pro-aktiv zu bearbeiten.

1 Hintergrund

Die Gesellschaft für Informatik e.V. (GI) ist mit rund 20.000 persönlichen und 250 korporativen Mitgliedern die größte und wichtigste Fachgesellschaft für Informatik im deutschsprachigen Raum. Mit 14 Fachbereichen, über 30 aktiven Regionalgruppen und unzähligen Fachgruppen ist die GI Plattform für alle Disziplinen in der Informatik. Weitere Informationen finden Sie unter www.gi.de.

Der Fachbereich Informatik in Recht und Öffentlicher Verwaltung der GI zielt auf Synergieeffekte durch die Verknüpfung der gewachsenen Kultur im öffentlichen Handeln mit der Sprach- und Denkwelt der Informatik ab. Insbesondere die Fachgruppe Rechtsinformatik versteht sich als interdisziplinäres Wissenschaftsgremium an der Schnittstelle zwischen Recht, Informatik, Ökonomie, Informationswissenschaft, Soziologie und Philosophie.

Die GI wurde vom Sachverständigenrat für Verbraucherfragen beauftragt, eine Machbarkeitsstudie für ein Algorithmengesetz im Bereich Verbraucherscoring zu erstellen. Der Titel der Studie wurde auf „Technische und rechtliche Betrachtungen algorithmischer Entscheidungsverfahren“ abgeändert, weil er den aktuellen wissenschaftlichen Diskurs und damit den Charakter des Gutachtens besser wiedergibt.

Folgende Experten wirken an dem Gutachten mit:

- ao. Univ.-Prof. Mag. Dr. Dr. Erich Schweighofer, Universität Wien (Sprecher und wissenschaftlicher Leiter)
- Prof. Dr.-Ing. Christoph Sorge, Universität des Saarlandes, juris-Stiftungsprofessur für Rechtsinformatik
- Prof. Dr. Georg Borges, Universität des Saarlandes, Lehrstuhl für Bürgerliches Recht, Rechtsinformatik, deutsches und internationales Wirtschaftsrecht sowie Rechtstheorie
- Prof. Burkhard Schäfer, The University of Edinburgh, Edinburgh Law School, Personal Chair of Computational Legal Theory
- Bernhard Waltl, M.Sc. M.A., Technische Universität München, Department of Informatics
- Dr. Matthias Grabmair, Carnegie Mellon University, School of Computer Science, Pittsburgh, PA, USA
- Daniel Krupka, Gesellschaft für Informatik e.V., Geschäftsführer (operative Leitung)

2 Einleitung

Dieses Gutachten „Technische und rechtliche Betrachtungen algorithmischer Entscheidungsfindungen“ wurde von der Gesellschaft für Informatik (GI), Fachgruppe Rechtsinformatik und der GI-Geschäftsstelle Berlin im Auftrag des Sachverständigenrates für Verbraucherfragen (SVRV), einem Beratungsgremium des Bundesministeriums der Justiz und für Verbraucherschutz (BMJV), erstellt.

In der heutigen Wissens- und Netzwerkgesellschaft ist Digitalisierung die dominierende Entwicklung. In dieser zunehmend digital vernetzten Welt wird der Einsatz der sogenannten Künstlichen Intelligenz (KI) / *Artificial Intelligence* (AI) zur Entscheidungsfindung eine immer größere Rolle spielen. In der öffentlichen Diskussion werden verstärkt Bedrohungsszenarien diskutiert. Im Kern der Betrachtung stehen *Big Data*, die intensive Sammlung von personenbezogenen Daten sowie deren Analyse und Einsatz zur Steuerung von verbraucherrelevantem Verhalten. *Big Data* bezeichnet Datenmengen, die so groß, komplex, schnelllebig und/oder schwach strukturiert sind, dass sie mit manuellen und herkömmlichen Methoden der Datenverarbeitung nicht ausgewertet werden können.

Effizienzerwartungen bedingen eine weitergehende Automatisierung und damit auch den Einsatz von Algorithmen zum Abschluss und zur Durchführung einer Vielzahl von Rechtsgeschäften, insbesondere auch mit Verbrauchern. Algorithmen werden zu zentralen Steuerungsmechanismen unserer Gesellschaft.[2]

2.1 Problemaufriss und Gegenstand der Studie

Im Kern dieser Studie steht die Kontrolle der algorithmischen Entscheidungsfindung (*Algorithmic Decision-Making* – ADM) aus technischer und juristischer Sicht. Beim ADM ist der Entscheidungsträger ein formaler Algorithmus – mit oder auch ohne Lernverfahren, bei dem ein Modell für das Verarbeiten von Daten (und gegebenenfalls anderen Algorithmen) trainiert wird. Es kommt daher auf die Auswahl der Daten und deren Bewertung an.

Rechtsfragen werfen sowohl die Auswahl der Daten, der Algorithmus selbst als auch die Lernverfahren auf. Lernende Algorithmen sind ein relativ neuer, aber faszinierender Regelungsgegenstand. Es geht um die Kontrolle dieser Instrumente intensiver Datenverarbeitung durch die Gesellschaft und den Rechtsstaat. Wenn wesentliche Entscheidungen in der „Glaskugel“ der algorithmischen Entscheidungsfindung (ADM) getroffen werden, besteht die Gefahr eines Ungleichgewichts zwischen den Vertragspartnern und – für den schwächeren Partner, den Verbraucher – die Gefahr der Hilflosigkeit (vgl. im Detail Kapitel 5.1.2).

Dies drückt sich im Titel eines Beitrages einer großen deutschen Tageszeitung aus: „Computer – darf ich raus?“[3] Hier wird, verbunden mit einem großflächigen Bild von Insassen eines US-amerikanischen Gefängnisses, die Nutzung algorithmischer Entscheidungen beschrieben. Derartige Beiträge in Massenmedien lassen sich wohl als Ausdruck eines Unbehagens gegenüber algorithmischen Entscheidungen verstehen.

[2] Martini 2017.
[3] Frankfurter Allgemeine Zeitung (17. März 2018) [https://bit.ly/2CDI3E1].

Dieses Unbehagen wird auch in der juristischen Diskussion formuliert,[4] insbesondere von Martini. Dieser stellt etwa die These auf, „Blackbox-Algorithmen" beschwörten „das dumpfe Gefühl herauf, überwacht zu werden" und könnten „die Anwandlung auslösen, nach undurchsichtigen Entscheidungskriterien diskriminiert zu werden oder zum Objekt sublimer Steuerung zu degenerieren".[5]

Unbehagen ist der Motor für Änderungswünsche bestehender rechtlicher Instrumente, aber für eine rechtliche Regelung zu ungenau. Vorab sind algorithmische Entscheidungen zu analysieren, um die juristischen Fragen besser zu kategorisieren (siehe Kapitel 5.1).

Eine umfassende Analyse und Systematisierung der Probleme algorithmischer Entscheidungen ist nicht Gegenstand dieser Studie. Die Untersuchung beschränkt sich auf Beurteilungen von Personen durch algorithmische Entscheidungen und damit verbundene zentrale Rechtsfragen. Ein wesentlicher Aspekt sind fehlerhafte algorithmische Entscheidungen, die aus rechtlicher Sicht unmittelbar problematisch sind, soweit dadurch Rechtsgüter Betroffener verletzt werden. Dabei sollen im Rahmen dieser Studie vor allem fehlerhafte Beurteilungen von Personen in ihrer Eigenschaft als Verbraucher betrachtet werden. Als verbraucherrelevant werden dabei solche Entscheidungen angesehen, die für Rechtsverhältnisse natürlicher Personen in ihrer Eigenschaft als Verbraucher von Bedeutung sind. Hier geht es insbesondere um die Existenz von Ungleichgewichtslagen durch den Einsatz von ADM sowie die Diskriminierung, d.h. die unerwünschte Ungleichbehandlung, der von den Entscheidungen Betroffenen anhand bestimmter Entscheidungskriterien in Algorithmen. Hier wird etwa herauszuarbeiten sein, welche Entscheidungskriterien möglicherweise als unzulässig gelten (offensichtlich etwa: Religionszugehörigkeit, sexuelle Orientierung etc. in Personalentscheidungen), wie die Verwendung solcher Entscheidungskriterien in Algorithmen eingebettet ist und wie eine die Kontrolle solcher ADM-Systeme erfolgen könnte (Kapitel 4). Ein anderes Gefährdungsszenario ist die inhaltliche Unrichtigkeit einer Entscheidung. Dabei ist etwa zu fragen, unter welchen Voraussetzungen eine maschinengenerierte Entscheidung als richtig anzusehen ist und wie die Richtigkeit von Entscheidungen durch Algorithmen zu adressieren ist (Kapitel 5).

Der Schwerpunkt der Studie ist somit die Analyse algorithmischer Entscheidungen aus technischer wie rechtlicher Sicht hinsichtlich dieser Fragestellungen. Als Grundlage der Bewertung dienen die Grund- und Menschenrechte sowie die ethischen Wertungen der Rechtsordnungen. Die technische Analyse beschreibt im Detail ADM und Maschinelles Lernen und versucht eine Kategorisierung der Probleme, insbesondere auch die Entstehung von Ungleichbehandlung, sowie deren Vermeidung durch „Fairness" im Maschinellen Lernen. Des Weiteren wird die Kontrolle von algorithmischen Entscheidungen als Ansatz zur Regulierung, wie z.B. Auditierung und Testing, behandelt. Rechtlich ergibt sich eine Notwendigkeit der Regulierung bei fehlerhaften Entscheidungen, wobei hier insbesondere das Diskriminierungsverbot eine wesentliche Rolle spielt.

Die ausgewählten Fälle werden vor dem Hintergrund vorher definierter Fälle mit Gefährdungsszenarien untersucht. Aus diesen werden die Rechtsfragen herausgearbeitet, die in der weiteren Analyse herangezogen werden.

[4] Siehe etwa Wischmeyer, AöR 2018, 1, 24, Fn. 92 mit zahlreichen Nachweisen.
[5] Martini 2017.

2.2 Beschreibung der Fragestellung in Fällen

In Übereinstimmung mit dem SVRV und unter Einbeziehung der Gefährdungsszenarien werden folgende Fälle von Verbraucherscoring identifiziert und besonders behandelt: Kreditscoring und Preisdifferenzierung. Der Schwerpunkt liegt aufgrund der Bedeutung beim Kreditscoring.

- **Kreditscoring:** Die Bewertung der Kreditwürdigkeit wird auf Basis einer statistischen Analyse ermittelt.
- **Preisdifferenzierung:** Anbieter fordern für die gleiche Leistung unterschiedliche Preise; Algorithmen bestimmen die zeitliche, räumliche, personelle oder sachliche Differenzierung.

Als Gefährdungsszenarien werden vom SVRV insbesondere folgende Sachverhalte identifiziert:[6]

- **„Inhaltlich unrichtiger Algorithmus“:** Unternehmen sammeln aus vielerlei Quellen Daten und erstellen ein Kreditscoring zur Entscheidungshilfe unter Einsatz von ADM-Verfahren. Die Identifikation von Parametern der Entscheidung und deren Bewertung im ADM sowie die Scores sind daher fehlerhaft, so dass die Ausfallwahrscheinlichkeiten für Kredite unrichtig sind.
- **„Diskriminierender Algorithmus“ (klassische Diskriminierungsmerkmale):** Das ADM knüpft an Parameter an, in denen eine direkte Diskriminierung oder auch indirekte Diskriminierung gesehen werden kann. Die Herausforderung liegt darin, dass die Datenlage ausreichend umfangreich sein muss, um festzustellen, dass der an sich nicht diskriminierende ADM doch zu einer Benachteiligung führt.
- **„Intransparent personalisierender Algorithmus“ (personalisierte Preise):** Eine Analyse der oft nur unvollständig vorliegenden Daten lässt vermuten, dass die verwendeten Entscheidungskriterien für die Preisgestaltung diskriminierende Elemente aufweisen, wenn nicht eine durch sachliche Kriterien gedeckte Preisfestsetzung vorliegt.

All diese Begriffe sind offen und die Überprüfung erfordert eine entsprechende Datenlage. Die Ungleichheit kann aber auch durch sachliche Kriterien gerechtfertigt sein. Der Transparenz kommt eine große Bedeutung zu. Dieser sind jedoch auch Grenzen gesetzt und sie stellt keine allgemeine Lösung dar, um das Problem diskriminierender Algorithmen rechtlich zu kontrollieren (Kapitel 7.4).

[6] Sachverständigenrat für Verbraucherfragen, Gefährdungslagen für Verbraucherinnen und Verbraucher durch ADM, 11. Dezember 2017.

Dies ergibt folgende Beschreibung in Fällen:

Ein Verbraucher erhält einen Kredit mit Hinweis auf sein Rating nicht. Die Bank verwendet ein ADM-System zur Bonitätsprüfung mit Angaben zu Alter, Geschlecht, Migrationshintergrund und Jahreseinkommen. Das ADM-System wertet a) Migrationshintergrund negativ und b) niedriges Jahreseinkommen negativ, wobei bekannt ist, dass dies insbesondere Frauen betrifft.

Ein Verbraucher muss einen höheren Preis zahlen, weil er a) kein Stammkunde ist, b) keiner sozial benachteiligten Gruppe angehört, c) als leistungsfähiger eingestuft wird, d) die Leistung in einer entlegenen Region erbracht wird, e) er einer Gruppe mit hohem Kreditausfallsrisiko angehört, f) aufgrund seines Namens als Migrant angesehen wird und g) die Dienstleistung sehr stark nachgefragt wird.

Eine wesentlich umfangreichere Beschreibung des Leitbeispiels „*Fair Lending*“ erfolgt in Kapitel 3.4.4.

Aus technischer Sicht steht die laborweise Prüfung der eingesetzten ADM-Verfahren auf mögliche Ungleichbehandlung im Vordergrund: Die ADM darf a) nicht unrichtig sein und b) weder mittelbar noch unmittelbar diskriminieren.

2.3 Vorgehen der Untersuchung

Im Hinblick auf die interdisziplinäre Kompetenz der Rechtsinformatik soll eine Brücke zwischen der rechtlichen und technischen Begrifflichkeit der Algorithmen geschlagen werden. Ziel der Studie ist eine sowohl interdisziplinäre als auch rechtsvergleichende Analyse der Möglichkeiten, die Gesetzgeber für die rechtliche Gestaltung der Entwicklungs- und Anwendungsumgebung von Algorithmen haben. Dies erfordert eine Analyse der Konzepte der Informatik mit der Begrifflichkeit des deutschen Rechts sowie der zum Vergleich herangezogenen Rechtsordnungen.

Juristisch geht es um richtige und faire Entscheidungen, die in der „analogen Welt“ durch Verhaltensvorschriften für Menschen sichergestellt werden sollten. Im digitalen Zeitalter werden Entscheidungen von Maschinen vorbereitet bzw. oftmals auch selbst getroffen. Ein wesentliches Merkmal ist die Massenhaftigkeit und Schnelligkeit der Entscheidungsprozesse bei ADM-Prozessen. Durch die Lernfähigkeit der Algorithmen können die Entscheidungskriterien laufend an die jeweiligen Zielsetzungen angepasst und verbessert werden. Der Entscheidungsraum mit den unzähligen Optionen kann durch den Einsatz von Algorithmen wesentlich besser analysiert und beherrscht werden. Die Entscheidungen werden durch die Algorithmen wesentlich gleichförmiger, was auch der standardisierten Eingabe geschuldet ist. Es liegt an der effektiven Kontrolle der Algorithmen, ob diese eine weitere wesentliche Verschlechterung der Verbraucherverträge bewirken können.

Es stellt sich die Frage, ob durch den Einsatz von Algorithmen der Entscheidungsprozess vollkommen neu gestaltet wird und daher neue rechtliche Regelungen möglich und notwendig sind.

Daher beginnt die Studie mit einer Analyse algorithmischer Entscheidungen aus technischer Sicht und mit einer Darstellung der Grundlagen von ADM und *Machine Learning* (siehe Kapitel 4). Dann wird hinterfragt, inwieweit die Entscheidungen von Algorithmen erklärt und einem Auditing unterzogen werden können. Dem Vorteil einer gleichförmigen Entscheidung steht der Nachteil einer „richtigen" rechtlichen und ethischen[7] Bewertung dieser quantitativen Kriterien und deren nötiger Anpassung in einem Rechtsstreit oder einer Entscheidung des Gesetzgebers gegenüber. Hierbei sind mögliche Parameter die Daten, das Training sowie die Auswahl der Attribute und des Modells.

Auszugehen ist jeweils von den verwendeten (personenbezogenen) Daten und deren Repräsentation als vieldimensionale Vektorräume. Es gibt eine Fülle von Methoden, mit denen Modelle und ihre Parameter automatisch „gelernt" werden können. Die Erklärbarkeit dieser Systeme (*Explainable AI*) bildet sowohl die Grundlage für die Kontrolle als auch für eine mögliche Auditierung bzw. Zertifizierung. Hierbei werden verschiedene Ansätze bewertet (ADM als Blackbox, ADM als Whitebox). Diese Überlegungen gehen in Richtung von Auditing-Verfahren, Testing bzw. eines „Algorithmen-TÜVs".

Bei der Betrachtung algorithmische Entscheidungen aus rechtlicher Sicht (Kapitel 5) sind Grundüberlegungen zu widerstreitenden Interessen und Schutzgütern vorzunehmen, um Kriterien für Diskriminierungen sowie fehlerhafte Entscheidungen festzustellen. Die bestehenden Instrumente der Nichtdiskriminierung sowie des Verbraucherschutzes werden analysiert.

Der internationale und rechtsvergleichende Teil der Studie (Kapitel 6) behandelt methodologische Vorbemerkungen zum Lernen von anderen Rechtsordnungen. Der Schwerpunkt wurde auf Großbritannien und die USA gelegt, weil beide Staaten Technologieführer im Bereich digitaler Technologien sowie ADM- und maschineller Lernverfahren sind und der Diskriminierungsschutz eine lange Tradition hat. Neben dem geschriebenen Recht (*law in books*) und der Rechtswirklichkeit (*law in action*) wird auch ein „Recht in Planung" berücksichtigt. Überlegungen zur Regulierung von ADM-Verfahren werden in vielen Staaten angestellt. Dabei müssen die Unterschiede zum jeweiligen Rechtssystem sehr genau berücksichtigt werden.

Die Analyse der bestehenden Rechtslage in Deutschland hinsichtlich unrichtiger bzw. diskriminierender algorithmischer Entscheidungen zeigt die Stärken und Schwächen bestehender Regeln und Instrumente (Kapitel 7).

Die Handlungsempfehlungen fokussieren Lehre/Forschung/Erkenntnis, die mögliche Institutionalisierung und den gesetzlichen Handlungsbedarf (Kapitel 8).

2.4 Struktur des Gutachtens

Der Bericht ist in acht Kapitel gegliedert, die jeweils aufeinander aufbauen bzw. sich ergänzen und zu der notwendigen ganzheitlichen Betrachtung (technologische und rechtliche Aspekte) beitragen.

[7] Vgl. dazu die Arbeiten der Bertelsmann Stiftung zur Algorithmenethik, zuletzt Rohde 2018.

- **Kapitel 2 „Einleitung“:** Zu Beginn wird die grundlegende Motivation für die Studie „Technische und rechtliche Betrachtungen algorithmischer Entscheidungsfindungen“ eruiert. Ausgehend von dem Befund, dass algorithmische Entscheidungsfindung mehr und mehr in gesellschaftlich relevante Bereiche vordringt, ist es naheliegend, dass sich die Politik mit diesem Thema insbesondere vor dem Hintergrund einer möglichen Diskriminierung beschäftigt.

- **Kapitel 3 „Terminologie und interdisziplinäre Methode“:** Das Thema ist ein genuin interdisziplinäres und bedarf der aktiven Zusammenarbeit und des strukturierten Austauschs zwischen der Informatik und der Rechtswissenschaft. Aus beiden Disziplinen werden Begriffe verwendet und eingeführt, die einer Erläuterung bedürfen. In diesem Kapitel werden die Begriffe erklärt und wird die Methode der interdisziplinären Zusammenarbeit beschrieben.

- **Kapitel 4 „Algorithmische Entscheidung aus technischer Sicht“:** Im Kern der Betrachtungen stehen Entscheidungen, die von Algorithmen getroffen werden. Es ist von zentraler Bedeutung, dass diese (Software-)Systeme detailliert beschrieben werden und das Problemfeld sorgfältig differenziert wird. Es handelt sich um eine große Menge verschiedenster Algorithmen, die genutzt werden können, und diese erfordern unter Umständen eine separate Betrachtung. Begriffe und Konzepte werden differenziert und mögliche technische Strategien, die zur Regulierung beitragen können, wie z.B. Auditierung und Testing, erläutert.

- **Kapitel 5 „Algorithmische Entscheidung aus rechtlicher Sicht“:** Auf Basis der technischen Sicht wird der rechtliche Rahmen der Entscheidungen von Algorithmen erläutert. Dabei wird der Fehlerbegriff differenziert nach Fehlern im Entscheidungsprozess (insbesondere Unzulässigkeit, Intransparenz, Grundlagen, Wertung), die einer separaten Betrachtung bedürfen. Davon wird der Begriff der Diskriminierung unterschieden und der Status quo der algorithmischen Entscheidung im deutschen Recht diskutiert.

- **Kapitel 6 „Regulierung und Standards auf internationaler Ebene“:** Im Rahmen des Zwischenberichts werden auch Regulierungen anderer Länder betrachtet. Von besonderer Bedeutung ist dies, weil die Algorithmen grundsätzlich und ohne weiteres in anderen Ländern eingesetzt werden können. Auch dort ist es naheliegend, dass sich die Politik dieser Herausforderung stellen muss, und eine Zusammenarbeit und ein Austausch erscheinen sinnvoll.

- **Kapitel 7 „Möglichkeiten der Regulierung in Deutschland“:** Aus der Synthese der Ergebnisse aus den einzelnen Kapiteln, die aufeinander aufbauen, sollen die Möglichkeiten aufgezeigt werden, die in Deutschland verfügbar sind bzw. deren Einsatz in Deutschland nicht auszuschließen ist.

- **Kapitel 8 „Handlungsempfehlungen“:** Den Abschluss der Betrachtungen des Zwischenberichts bilden Handlungsempfehlungen für die mögliche Gestaltung eines Algorithmengesetzes. Dabei sollen auf Basis der vorgehenden Überlegungen die Möglichkeiten, aber auch die Probleme und Herausforderungen berücksichtigt werden.

3 Terminologie und interdisziplinäre Methode

3.1 Grundlegende Begriffe und Definitionen

Für ein besseres Verständnis sollen zu Beginn einige Begrifflichkeiten definiert werden:

Entscheidungsfindung (Decision-Making): Entscheidungsfindung beschreibt einen Prozess, bei dem ein Entscheidungsträger eine Aktion aus mehreren Alternativen auf Basis qualitativer und quantitativer Attribute auswählt.

Algorithmische Entscheidungsfindung (Algorithmic Decision-Making – ADM): Als ADM bezeichnet man eine Entscheidungsfindung dann, wenn der Entscheidungsträger ein Algorithmus ist.

Erklärung: Eine Erklärung beschreibt den logischen und kausalen Zusammenhang zwischen den Gründen und einer ausgewählten Aktion in einer festgelegten Sprache.

Attribut (auch **Feature):** Ein Attribut, oder Feature, ist eine beobachtbare und messbare Eigenschaft einer realen Entität, z.B. einer Person oder eines Objekts.[8]

Algorithmus: als Computerprogramm implementierbarer Ablauf von Handlungsanweisungen zur Lösung eines formal definierten bzw. definierbaren Problems.

Lernender Algorithmus: ein Algorithmus, der durch das Verarbeiten von Daten (und gegebenenfalls anderen Algorithmen) ein Modell trainiert.

Maschinelles Lernen: Forschung und Anwendung von Algorithmen, die eine bestimmte Aufgabe bewältigen und ihre Leistung/Performanz durch eine Form von Erfahrung (z.B. Trainingsdaten mit „korrekten Lösungen“) verbessern.

Modell: Kombination von Algorithmen mit Daten und/oder trainierten Parametern, die auf der Basis eingegebener Daten neue Informationen produziert (z.B. Wertvorhersage).

Training: Prozess der Parameteroptimierung für ein Modell anhand von Trainingsdaten und Optimalitätskriterien.

Test: die Evaluierung eines Modells auf Basis von Testdaten.

3.2 Interdisziplinäre Methoden – Recht trifft Informatik

Die rechtliche und ethische Bewertung von Entscheidungsprozessen auf Richtigkeit und Nichtdiskriminierung ist eine wesentliche Aufgabe der Rechtswissenschaft. Der Einsatz von ADM wirft neue Fragen auf. So gibt es zahlreiche Varianten von algorithmischen Entscheidungsverfahren; dabei gilt es zu untersuchen, welche der Unterschiede rechtlich relevant sind. Hier ist zu berücksichtigen, dass ADM oft einem Unbehagen aufgrund potenziell undurchsichtiger Entscheidungskriterien und möglicher Überwachung begegnen.

[8] Bishop 2006.

Eine interdisziplinäre Zusammenarbeit ist erforderlich, um wechselseitig die neuen Herausforderungen zu verstehen und Lösungen auszuarbeiten.

Idealerweise besteht ein gemeinsames Ziel zwischen Juristen und Technikern: die optimierte und rechtskonforme Konzeption und Implementierung von ADM. Daher muss das Problembewusstsein gestärkt und verbessert werden. Algorithmen eröffnen neue Chancen und Herausforderungen für die Entwicklung von rechtlichen Regimen zur Governance von Entscheidungsprozessen. Der Einsatz von Algorithmen kann die Ambiguität verringern, die Verantwortlichkeit gegenüber der Öffentlichkeit erhöhen und letztlich mehr Transparenz in einen früher geheimen Entscheidungsprozess bringen.

Der Jurist hat mit rechtsdogmatischen Methoden den Sinn des Gesetzes zu erforschen und auszulegen; hier stehen Fehler im Entscheidungsprozess im Vordergrund. Es gibt zum Teil Vorschriften für die Entscheidungsfindung – die jeweiligen Verfahrensrechte –, aber diese sind auf Input, Bewertung und Entscheidungskriterien konzentriert. Kennzeichen ist auch ein oft relativ großer Ermessensspielraum in der Interpretation der Begriffe.

Algorithmische Entscheidungsverfahren erfordern eine Determinierung des Entscheidungsprozesses. Bei der Prüfung auf Fehler und Diskriminierung muss dieser – soweit möglich – detailliert betrachtet werden. Der Erklärbarkeit von Algorithmen kommt damit eine besondere Bedeutung zu.

Algorithmen sind in erster Linie ein Mittel zur Entscheidungsfindung; in einigen Fällen auch eine direkte Anwendung von Rechtsvorschriften. Standards und Zertifizierungen sind erst am Anfang der Entwicklung. Die prozedurale Prüfung der individuellen Entscheidung stand bisher im Vordergrund; beim Einsatz von Algorithmen kommt der Prüfung der gesamten Entscheidungspraxis und deren Fehler bei bestmöglicher Erklärbarkeit eine stärkere Bedeutung zu.

Bei Informatikern steht die technische Entwicklung von ADM im Vordergrund. Hier sollte eine ausformulierte Spezifikation des Systemverhaltens in allen Situationen gegeben sein. Die eingesetzten Algorithmen und Trainingsverfahren können durch Auditierung und Testen auf Fairness überprüft werden. Für einen diskriminierungsfreien und fehlerfreien Einsatz von ADM müssen hier die rechtlichen Vorgaben effektiv einfließen.

Methodik und Terminologie sind und bleiben weitgehend unterschiedlich; sie können aber in einer Kooperation zusammengeführt werden. Nur durch laufende Interaktion kann ausreichend Verständnis für die jeweilige Methodik gewonnen und diese angepasst und optimiert werden.

Beim Ziel einer gerechten Entscheidung wird häufig auf unbestimmte Begriffe wie Fairness abgestellt, für die es aber vorab keine umfassende Beschreibung oder Spezifikation gibt. Juristisch ist die Begriffsklärung erst bei stabiler Rechtsprechung ausreichend. In der Informatik muss diese Begriffsklärung aber schon früher viel detaillierter abgebildet werden. Die Lösung dieser Herausforderung der Ambiguität ist entscheidend für die Qualität des Dialogs. Juristen sind diese Ambiguität gewohnt, bei ADM mit statischen und Machine-Learning-Verfahren fließt diese auch in den Algorithmus mit ein. Informatiker müssen lernen, dieses Ermessen im Sinne der Prinzipien bestmöglich zu nutzen, und sollten hierbei mit Juristen zusammenarbeiten.

Juristen und Informatiker müssen daher versuchen, einander wechselseitig besser zu verstehen und mehr zusammenzuarbeiten, um die Herausforderungen eines präzisen Algorithmus und einer unscharfen Rechtssprache besser zu bewältigen. Weitere

Unbestimmtheiten liegen in der Unsicherheit über die Anwendungsfälle, die Auswirkung von bestimmten Parametern oder der Problemlösung. Die Entwicklung einer gemeinsamen Ontologie könnte hilfreich sein.

Für die Erstellung, Implementierung und Evaluierung sollten Teams von Juristen und Informatikern gebildet werden, die auch die Details des notwendigen Wissens beherrschen und durch wechselseitige und laufende Zusammenarbeit ausreichend Kommunikationsfähigkeit für die jeweils andere Methode bzw. Problemstellung haben. Diese sollte durch Aus- und Weiterbildung laufend verbessert werden.

Der Einsatz von Spezialisten wie Data Scientists könnte hilfreich sein, weil diese mit ihrer Spezialisierung die Fragestellungen beider Disziplinen abdecken und in weiterer Folge als Vermittler hinsichtlich der Terminologie bzw. Methodik fungieren können.

Die verstärkte Zusammenarbeit erfordert Maßnahmen in der Ausbildung von Experten sowie auch in der Weiterbildung von Juristen und Informatikern (siehe 8.1).

Daher liegt es nahe, hier ein **ständiges Beratungsgremium** einzurichten, in dem Anwendungsprobleme wie auch Forschungsfragen diskutiert und hinterfragt werden können.

3.3 Explorativ-empirische Untersuchung der Praxis um ADM

Um den aktuellen Stand des Einsatzes von ADM-Systemen in der Praxis zu untersuchen und Einschätzungen zur Machbarkeit bestimmter Regulierungsansätze zu erhalten, wurde im Laufe der Erstellung des Gutachtens eine Serie von Gesprächen mit Praktikern in diesem Bereich geführt. Die so gewonnenen Erkenntnisse förderten die interne Diskussion und werden im Gutachten an verschiedenen Stellen explizit aufgegriffen.

Die Untersuchung umfasste Hintergrundgespräche mit folgenden Organisationen/Personen:

- **Dr. Matthias Knecht**, Gründer und Geschäftsführer des Factoring-Startups **Billie.io,** das durch ADM-Technologie unterstützt Liquiditätsprüfungen von Firmen vornimmt (Gesprächsdatum: 3. Mai 2018)
- **Serena Holm**, Bereichsleiterin Corporate Affairs, und **Christopher Scheel**, Senior-Experte Public Affairs der **Schufa Holding AG**, welche unter anderem Kreditscoring von Verbrauchern anbietet (22. Mai 2018)
- **Prof. Dr. Robert Klein**, Inhaber des Lehrstuhls für Analytics & Optimization an der Wirtschaftswissenschaftlichen Fakultät der **Universität Augsburg**, der sich intensiv mit Fragen der Preisfindung auseinandersetzt (24. Mai 2018)
- **Ulf Linke**, Referatsleiter Grundsatzfragen Verbraucherschutz und **Astrid Gruschka**, Referatsleitung Kompetenzzentrum Verbraucherschutz Banken, die sich bei der **Bundesanstalt für Finanzdienstleistungsaufsicht (BaFin)** mit Fragen des Verbraucherschutzes im Rahmen der Finanzmarktaufsicht beschäftigen (11. Juni 2018)
- **Michael Kaiser**, Leiter Referat 3.2 (unter anderem Banken, Sparkassen, Kreditinstitute, Auskunfteien (z.B. Schufa), Inkasso) beim **Hessischen Beauftragten**

für Datenschutz und Informationsfreiheit, der unter anderem die Schufa beaufsichtigt (09. Juli 2018)

Die Gespräche wurden jeweils vom Projektleiter (Daniel Krupka) sowie mindestens einem der Fachautoren geführt. Die typische Gesprächsdauer betrug ungefähr eine Stunde.[9] Nach einer anfänglichen Vorstellungsrunde wurden das Gebiet und das Ziel der Studie erklärt. Die Gesprächsführung war weitestgehend frei, orientierte sich jedoch an den folgenden Kernfragen:

- Welche ADM-Technologien kommen in Ihrer Organisation bzw. Ihren jeweiligen Branchen zum Einsatz?
- Welches Problembewusstsein und welche Lösungsansätze existieren in Ihrer Organisation bzw. Branche zum Thema Diskriminierung durch ADM-Verfahren?
- Welche Regulierungsmöglichkeiten von ADM halten Sie für nützlich, wünschenswert, durchsetzbar oder ungeeignet?

3.4 Internationale und rechtsvergleichende Methoden

Die Betrachtungen dieser Studie sollen in einen internationalen und rechtsvergleichenden Kontext gestellt werden. Das Ziel ist zum einen, aus der Erfahrung anderer Rechtsordnungen zu lernen, und zum anderen sicherzustellen, dass keiner der Vorschläge, die in dieser Untersuchung gemacht werden, internationale Harmonisierungsbemühungen vorwegnimmt oder mit internationalen Aktivitäten in Konflikt steht.

Insbesondere das Lernen von anderen Rechtsordnungen ist nicht unproblematisch. Die Transplantation von erfolgreichen Lösungen aus einer juristischen Tradition in eine andere kann unvorhergesehene Konsequenzen haben, wenn der „Fremdkörper“ mit juristischen Begriffen und Vorstellungen der empfangenden Rechtsordnung interagiert.[10] Dies hat in Teilen der Literatur zu einer generellen Skepsis gegenüber dieser Art des rechtsvergleichenden Lernens geführt, insbesondere wenn der Transfer der Konzepte die Grenzen zwischen den großen Rechtstraditionen (etwa vom *Common Law* zum kontinentaleuropäischen *Civil Law*) überschreitet.[11] Andere Stimmen sind optimistischer, betonen aber trotzdem, dass Rechtsordnungen komplexe holistische Einheiten sind, die es nicht zulassen, einzelne Begriffe isoliert zu betrachten, und dass der weitere rechtsbegriffliche Kontext sowie auch das politische, ökonomische und historische Umfeld mit in die Betrachtung einbezogen werden müssen.[12]

Um die Aussagekraft des rechtsvergleichenden Teils der Studie zu erhöhen und dem SVRV Hinweise zur Vergleichbarkeit und damit Reichweite der rechtsvergleichenden Ergebnisse zu geben, werden im Folgenden einige dieser Arbeit zugrundeliegenden Annahmen explizit gemacht und wird die Motivation für die methodologischen Entscheidungen erklärt. Das ist vor allem deshalb notwendig, da rechtsvergleichende Analysen an der Schnittstelle zwischen Technologie und Recht neue methodologische Probleme aufwerfen, für die es

[9] Die Ausnahme war ein kürzeres Gespräch mit der Wertpapiergruppe der BaFin, aus dem sich die Planung eines weiteren Gesprächs mit der Kreditvergabeaufsicht ergab.
[10] Teubner 1998.
[11] Legrand 1996.
[12] Watson 1996; Miller 2003; Levi-Faur 2005.

noch keine allgemein akzeptierten Lösungen gibt – obgleich diese Art der Analyse bei politischen Entscheidungsträgern immer beliebter wird.[13]

3.4.1 Methodenpluralismus

Ausgangspunkt unserer Studie ist ein funktionalistischer Ansatz in der Tradition von Zweigert und Kötz, den diese auch gerade für den grenzüberschreitenden Transfer von Konzepten geeignet hielten. Die Gefährdungsszenarien, die diese Analyse im ersten Teil der rechtlichen Fragestellung leiten, werden hier als *Tertium Comparationis* zwischen den Rechtsordnungen verwendet. Die Annahme dabei ist, dass zumindest ökonomisch ähnliche Gesellschaften auch mit ähnlichen „lebensweltlichen" Problemen konfrontiert sind: Die Probleme sind fixiert, die rechtlichen Lösungen variieren und befinden sich im Fluss.

Diese Analyse zeigt, dass dieses Modell zwar einen guten Ausgangspunkt bildet, aber nicht ohne Weiteres direkt angewendet werden kann: Wissenschaft und Informationstechnologien blicken ihrerseits auf Recht und Politik und wollen von dort einerseits Hinweise dazu, welche Forschung in rechtskonforme Produkte umgesetzt werden kann, und andererseits die Definition von Problemen, für die sie dann Anwendungen und Lösungen entwickeln. Dies bedeutet zum einen, dass die „lebensweltlichen" Probleme nicht, wie Zweigert und Kötz glaubten, unabhängig von Fragen des Rechts existieren. Zum anderen bedeutet es, dass es gerade in jungen Forschungsgebieten und neuen Technologien unterschiedliche Ansätze mit divergierender Nomenklatur, Philosophie und Methoden gibt, die mehr oder weniger mit der Jurisdiktion korrespondieren. Im Bereich des Technologierechts ist dieses methodische Problem besonders spürbar. Da zunehmend unter dem Motto „*Code is Law*" die Implementierung rechtlicher Vorschriften direkt in der Computersprache propagiert wird, um damit rechtskonformes Verhalten der Technologie durch Design zu gewährleisten, wird immer mehr und sehr wörtlich die Lebenswelt durch Recht „durchtränkt" – so spricht etwa Hildebrandt von *ambient law*.[14] Im Bereich dieser Studie sehen wir dies z.B. in der Entwicklung von *interpretable AI* oder *explainable AI*, in der traditionelle Ziele der KI-Forschung auch gerade durch die sich ändernde regulatorische Landschaft stark beeinflusst werden und Forscher aus unterschiedlichen Rechtsordnungen unterschiedliche Schwerpunkte setzen.[15] Die Auswirkungen dieser Integration von Recht direkt in Technologie für die rechtsvergleichende Methodik sind noch nicht aufgeholt, wobei aber gerade für Funktionalisten die Probleme offensichtlich sein sollten.

Science, Technology and Society Studies (STS) verwenden soziologische Ansätze, um die Entstehung von Technologie-Communitys und ihrer Nomenklatur zu analysieren.[16] In der Rechtsvergleichung entspricht dieser Denkansatz dem *Comparative Law and Culture*-Ansatz (CLC), der in bewusster Abgrenzung zum Funktionalismus entwickelt wurde. In dieser Arbeit werden beide Methoden verbunden, ohne dabei den „transferkritischen" Ansatz des CLC zu übernehmen. Stattdessen werden das weitere soziale, politische und kulturelle Umfeld als „Regulativ" verwendet, um aufzuzeigen, wo trotz scheinbarer Ähnlichkeit in der Problemstellung und dem technischen Vokabular tieferliegende Unterschiede vorhanden

[13] So hat z.B. die britische Regierung Studien zur Kontrolle von ADM-Systemen beauftragt, die auch gerade deutsche und andere kontinentaleuropäische Ansätze berücksichtigen sollen.
[14] Hildebrandt 2008.
[15] Der erste internationale Workshops zu diesem Thema, die *International Joint Conference on Artificial Intelligence: Workshop on Explainable Artificial Intelligence (XAI)*, im Jahr 2017 unterstreicht diese Entwicklung.
[16] Williams/Edge 1996; Bauchspies et al. 2005.

sein könnten, die eine direkte Übertragung der Erfahrung aus anderen Rechtsordnungen ins deutsche Recht wenn nicht unmöglich, so doch risikoreicher machen.

Da es eine zentrale Frage für den Gesetzgeber ist, wie sich verschiedene Regulierungsansätze wirtschaftlich auswirken werden, wäre zudem eine Analyse aus der Perspektive von „*Comparative Law and Economics*“[17] im Prinzip wünschenswert. Eine erste Literaturanalyse zeigte aber, dass es die dafür notwendigen international vergleichbaren Datensätze nicht gibt und diese auch im Rahmen einer Kurzstudie wie der vorliegenden nicht systematisch entwickelt werden können. Der einzige internationale Index von der Art, wie er für eine solche Analyse benötigt würde, ist der internationale „*Government AI Readiness Index*“.[18] Dieser analysiert indes nur, wie gut Regierungen aufgestellt sind, um Künstliche Intelligenz anwenden zu können, nicht aber, wie weit sie damit sind, diese Anwendungen zu regulieren. Etwas anders sieht die Situation für „traditionelle“ algorithmische Verfahren aus, also Verfahren, die nicht spezifisch Maschinelles Lernen oder KI im heutigen Sinne verwenden. Hier gibt es Studien zur Auswirkung von Regulierung insbesondere in der Kreditindustrie. Die zu erarbeitende Studie verwendet diesen Ansatz daher nur indirekt und führt keine systematische Analyse der Ergebnisse durch die *„Comparative Law and Economics“*-Perspektive durch, offenbart aber womöglich, wie effizient oder ineffizient in der Vergangenheit die Regulierung ähnlicher Probleme war.

3.4.2 Auswahl der Rechtsordnungen

Die einführende Diskussion zeigt, dass Lernen von anderen Rechtsordnungen dann am erfolgversprechendsten ist, wenn eng verwandte Systeme trotz alledem sehr unterschiedliche Lösungen entwickeln und mit ihnen Erfahrung sammeln können.

Eine erste Übersichtsanalyse der Literatur, die durch eine Präsentation im Rahmen einer kontinentaleuropäischen Konferenz (Internationales Rechtsinformatik Symposion IRIS 2018) validiert wurde, belegt, dass bislang noch keine umgesetzten Beispiele für lernende Algorithmen-spezifische Regelungen im Verbraucherschutzrecht bestehen, aus deren Erfahrung gelernt werden könnte.[19] Außerhalb des europäischen Datenschutzrechts gibt es erste Schritte im Finanzmarktrecht, so insbesondere die MiFID-II-Vorschriften zum algorithmischen Hochfrequenzhandeln, die in Deutschland weitgehend durch das *Gesetz zur Vermeidung von Gefahren und Missbräuchen im Hochfrequenzhandel* vorweggenommen wurden. Die sehr unterschiedliche Regelungsmaterie macht direkte Vergleiche schwierig, doch lassen sich zumindest Rückschlüsse zur technischen Machbarkeit von Archivierung großer Mengen sequenzieller Daten ziehen, die Teil eines Prüfungsregimes sein müssen.

Trotz des weitgehenden Fehlens einschlägiger Gesetze wurde versucht, Beispiele zu diskutieren, die die unterschiedlichen Rechtstraditionen in besonders typischer Weise vertreten. Frankreich repräsentiert die zivilistische Tradition mit dem *Gesetz zu einer Digitalen Republik*. Dies ist ein Versuch, außerhalb des Datenschutzrechts einen Anspruch auf transparente und faire Algorithmen in einem Verbraucherschutzkontext zu schaffen. Allerdings ist der Kern des neuen Rechts bestenfalls tangential zu den hier diskutierten Fallszenarien. Auch gibt es bislang keine Erfahrung mit der Umsetzung und gerichtlichen Auslegung dieses Rechts. Trotzdem stellt es als einer der seltenen Vorschläge der Kodifizierung eines Algorithmenrechts eine besonders typische „zivilistische“ Lösung dar.

[17] Siems 2005.
[18] Oxford Insights: Government AI Readiness Index, zuletzt besucht am 26.07.2018 [https://bit.ly/2ynQDiL].
[19] Siehe z.B. Pasquale 2015; Pasquale 2017.

Skandinavien wird durch einen aktuellen Fall zum algorithmischen Kreditscoring repräsentiert. Unsere Literatursuche deutete an, dass sich die rechtliche Diskussion in Kanada, Südamerika, Neuseeland[20], Australien[21], Asien und Afrika, soweit sie überhaupt stattfindet, weitgehend an amerikanischen oder EU-Vorschlägen orientiert.[22] Die Mehrheit der Vorschläge, die wir diskutieren werden, kommt daher aus Großbritannien[23] und den USA. Obgleich dies eine Einschränkung bedeutet, erlaubt es doch das erwünschte vergleichende Lernen und stellt eine rechtfertigbare Auswahl dar. Dies zum einen, weil die beiden Staaten in der Entwicklung der relevanten Technologie führend sind – der *Government AI Readiness Index* listet sie an erster und zweiter Stelle auf und hat diese beiden Länder als besonders einflussreich identifiziert. Dies hat zu besonders großem Regelungsbedarf geführt, der in aktive und gut dokumentierte Diskussionen des Gesetzgebers mündete. Zum anderen decken die USA und Großbritannien relevante Permutationen ab: Die USA haben, wie auch Deutschland, Aspekten des Gleichbehandlungsgebots Verfassungsrang gegeben. Dies ist in Großbritannien nicht der Fall, wo die ungeschriebene Verfassung eine derartige Verankerung nicht zulässt. Andererseits hat Großbritannien durch den *Equality Act 2000* die relevanten EU-Direktiven in das nationale Recht überführt und damit einen Rechtsrahmen geschaffen, der dem deutschen in dieser Hinsicht äquivalent ist. Wie das Land nach dem Ausscheiden aus der Europäischen Union damit verfährt, bleibt abzuwarten. Zumindest derzeit gilt in Großbritannien auch das relevante europäische Verbraucherschutzrecht, während die USA in diesem Bereich eine historisch und systematisch sehr unterschiedliche Entwicklung erfahren haben.

Auch das Datenschutzrecht wird in vielen Rechtsordnungen als eine Regulierungsmöglichkeit für KI diskutiert.[24] Hier unterscheidet sich der „sektorspezifische" Ansatz in den USA besonders deutlich von der EU. Innerhalb der Europäischen Union hingegen hat Großbritannien immer schon eine „minimalistische" Position vertreten, die die Bedeutung des Datenschutzrechts so weit als möglich zurückstuft. Dies ist zum einen ein Ergebnis des rechtlichen Kontexts (aktive Zurückweisung eines allgemeinen *privacy tort*

[20] Siehe Sek. IV des Berichts der New Zealand Human Rights Commission 2018.
[21] Siehe die Diskussion in dem Productivity Commission Draft Report: Data Availability and Use 2016, der europäisches Datenschutzrecht als Antwort auf algorithmische Diskriminierung empfiehlt.
[22] Siehe dazu auch den Algorithmenreport der WWW-Foundation [https://bit.ly/2O50iDC], der gezielt Beispiele außerhalb der EU/US-Debatte gesucht hat. Auch die Studie von Kathleen Siminyu „Artificial Intelligence in Low/Middle Income Countries, The East African Experience" [https://bit.ly/2wYtuER] weist nicht auf spezifisch rechtliche Diskussionen hin – obgleich in dem jungen afrikanischen AI-Umfeld Lösungen entwickelt werden, die einerseits von denen in Europa und den US abweichen, auch bedingt durch die oft unzuverlässige IT-Infrastruktur, aber ähnliche Probleme aufwerfen sollten. So beschreibt sie etwa Tala, eine Stiftung, die Mikrokredite in Kenia und Tansania anbietet. Bedingung ist, dass der Kunde eine Smartphone-App benutzt, die biographische Daten sammelt, aber auch die Kontakte der Lohnbewerber, die Größe ihrer Netzwerke und Unterstützungssysteme, Mobilität und Routineverhalten wie den täglichen Anruf bei den Eltern, um die Kreditwürdigkeit zu ermitteln.
[23] „Großbritannien" und „Britisches Recht" werden im Folgenden als Abkürzung für das Recht Englands, Schottlands und Nordirlands verwendet, da die für die Studie relevanten Gesetze typischerweise entweder Bundesrecht sind oder, wo sie *devolved matters* betreffen, nicht zu signifikanten Abweichungen zwischen den Regionen geführt haben. Es gibt bislang auch keine Initiativen der Regierungen in Belfast oder Edinburgh, eigene Algorithmengesetze einzuführen. Für die USA werden Unterschiede zwischen Bundes- und Landesrecht expliziter diskutiert werden müssen.
[24] Goodman/Flaxman 2017; Tene/Polonetsky 2012; Ishii 2017; Citron/Pasquale 2014; Wang/Wang 2017; Joshi 2018; Thelisson et al. 2017.

durch die Gerichte im 19. und frühen 20. Jahrhundert), zum anderen der kulturell-politischen Erfahrung geschuldet (keine totalitären Regime seit Cromwell im 17. Jahrhundert). Großbritannien und die USA eignen sich daher besonders für eine Triangulierung der *Common-Law*-Diskussion in dem für diese Studie relevanten Bereich, mit dem britischen Recht als mögliche „Brücke" zwischen kontinentaleuropäischen und US-amerikanischen Regulierungsphilosophien.

3.4.3 Verwendete Materialien

Rechtsvergleichende Forschung unterscheidet häufig zwischen dem geschriebenen Recht *(„law in books")* und dem praktizierten Recht *(„law in action")*, um systematische Unterschiede zwischen rechtlichem Ideal und praktischer Umsetzung zu verdeutlichen und zu evaluieren. Im Kontext dieser Studie ist das insbesondere für die unterschiedlichen Erfahrungen mit der Durchsetzung verbraucherschutzrechtlicher Normen relevant. Wie wir im Laufe dieser Betrachtung sehen werden, gibt es bislang keine Erfahrung mit neuen, speziell für die Herausforderung durch algorithmische Entscheidungen entwickelten Gesetzen. Die wenigen erfolgreichen Gesetzesinitiativen sind zu jung, um ihre Auswirkung auf wirtschaftliche Praxis und rechtliche Durchsetzung evaluieren zu können. Etwas fruchtbarer ist das Fallrecht zu diskriminierender Anwendung von Algorithmen unter etablierten rechtlichen Kriterien, und damit die analoge Anwendung von Regeln zur menschlichen Entscheidungsfindung auf Computersysteme. Naturgemäß diskutieren diese Entscheidungen aber häufig gerade nicht im Detail die spezifischen neuen Fragestellungen und Probleme der Anwendung von *Machine Learning*. Trotzdem erlauben sie uns zumindest, einige rechtsordnungsübergreifende Problemkonfigurationen zu identifizieren, und die Hypothese, dass die neue Technologie unzureichend reguliert ist, zu konkretisieren und zu verfeinern.

Um aber die empirische Grundlage zu verstärken, haben wir diese Kategorien um eine weitere erweitert: „Recht in Planung". Das heißt, zusätzlich zu formalen rechtlichen Quellen und Studien zu ihrer Umsetzung im Rechtsalltag ziehen wir auch offizielle Diskussionsdokumente, Vorschläge und akademische Studien hinzu, sofern diese von den zuständigen Gesetzgebern entweder direkt beauftragt wurden oder durch institutionelle Prozesse wie offizielle Anhörungen etc. besonders nahe an der Umsetzung sind.

Dieses Vorgehen soll sicherstellen, dass das Ziel einer rechtsvergleichenden und internationalen Studie – Kompatibilität mit internationalen Entwicklungen – erreicht wird. Rein abstrakte Diskussionsbeiträge von Akademikern aus anderen Rechtsordnungen, die bislang nur von der Wissenschafts-Community rezipiert werden, haben wir hingegen soweit relevant in den anderen Teilen dieser Untersuchung berücksichtigt.

3.4.4 Leitbeispiel Kreditscoring: Regelung, Aufsicht und Praxis im Bereich „Fair Lending" in den USA

Das Gutachten benutzt an vielerlei Stellen Beispiele aus dem Bereich der Kreditwürdigkeitsprüfungen. Um die Probleme und Lösungsansätze im Zwischenspiel von Technik/Statistik und Recht besser veranschaulichen zu können, gehen wir an dieser Stelle kurz auf ein breit dokumentiertes Fallbeispiel aus den USA ein. Dort regelt der *Equal Credit Opportunity Act*, spezifisch *Regulation B*[25], dass die Kreditvergabe nicht auf der Grundlage

[25] Federal Register, Bureau of Consumer Financial Protection, 21.12.2011, zuletzt online gesehen am 08.08.2018 [https://bit.ly/2N1ViD9]

von geschützten Attributen[26] des Antragstellers erfolgen darf. Zur Aufsicht dieses Verbots werden regelmäßig sogenannte *Fair Lending Examinations* durchgeführt. Zuständige Behörde ist hierfür das zum Treasury Department gehörende *Office of the Comptroller of the Currency* (OOC).[27] Das OOC veröffentlicht Richtlinien zur Aufsicht von Banken bei der Kreditvergabe[28] und wendet diese auch selbst an. Dort wird unterschieden zwischen einer ungleichen Vergabepraxis (*Disparate Treatment*) und einer vermeintlich neutralen Vergabepraxis, die aber auf gleich zu behandelnde Gruppen verschiedene Auswirkungen hat (*Disparate Impact*).

Die erste Gruppe umfasst zum einen offene Ungleichbehandlung (*Overt Evidence of Disparate Treatment*) nach geschützten Merkmalen in den kommunizierten Entscheidungskriterien der Bank wie beispielsweise unterschiedliche interne Richtlinien zur Kreditvergabe an Personen mit verschiedenen ethnischen Hintergründen. Zum anderen kann Ungleichbehandlung durch eine vergleichende Betrachtungsweise ermittelt werden (*Comparative Evidence of Disparate Treatment*). Wenn eine Bank zwei nach einem geschützten Merkmal verschiedene, aber sonst vergleichbare Antragsteller ungleich behandelt, muss sie eine Rechtfertigung liefern, welche von der Aufsicht geprüft wird. Prüfungsmaßstab ist konsequenterweise hierbei die Verneinung der „sonstigen Vergleichbarkeit".

Die zweite Gruppe betrifft Fälle, in denen eine Regelung angewendet wird, die zwar geschützten Attributen gegenüber prinzipiell neutral ist, aber faktisch Mitglieder bestimmter Gruppen negativer oder positiver behandelt als Mitglieder anderer Gruppen. Hierbei wird von „unausgewogenen Effekten" gesprochen (*disparate impact* oder *effects test*).[29] Auf rechtswidriges Verhalten wird hier erst nach Prüfung eines legitimen Handlungszwecks und der Verhältnismäßigkeit geschlossen. Bei solchen Prüfungen können insbesondere sogenannte *Regressionsverfahren* zur Anwendung kommen, wie wir sie in Kapitel 4.1 ausführlicher erläutern werden.

Eine Prüfung der Kreditvergabepraxis einer Bank erfolgt zweistufig.[30] In Stufe eins werden Bankdaten herangezogen, die zur Erfüllung der Pflichten des *Home Mortgage Disclosure Acts* (HMDA) erhoben wurden. Hieraus wird zunächst ein Prüfungsdatensatz erstellt, indem aus Vergabedaten bestimmter Kreditprodukte Gruppen von vergleichbaren Anträgen gesammelt werden. Hierbei handelt es sich um Gruppen von Anträgen einer zu schützenden Minderheitengruppe sowie deren Vergleichsgruppen, die von Mitgliedern der Mehrheitsgruppe gestellt wurden, aber im Übrigen äquivalent sind. Die Entscheidungen über die Anträge in den Vergleichsgruppen werden dann statistisch verglichen. Wenn die Unterschiede signifikant genug sind, kann die Prüfbehörde eine detaillierte Regressionsanalyse veranlassen (Stufe zwei). Es werden in einem leicht abgewandelten Prozess abermals solche Paardatensätze erstellt und die einzelnen Antragsdaten

[26] Die Regelung definiert eine „verbotene Entscheidungsgrundlage" wie folgt: *„Prohibited basis means race, color, religion, national origin, sex, marital status, or age (provided that the applicant has the capacity to enter into a binding contract); the fact that all or part of the applicant's income derives from any public assistance program; or the fact that the applicant has in good faith exercised any right under the Consumer Credit Protection Act or any state law upon which an exemption has been granted by the Bureau."*

[27] Offizielle Webseite [www.occ.treas.gov/]; siehe auch Wikipedia [https://bit.ly/2MYCQva].

[28] OOC Fair Lending Handbook (Version of January 2010) [https://bit.ly/2N2Qlde].

[29] Im deutschen Recht entspricht dies der Kategorie der mittelbaren Benachteiligung.

[30] Die Autoren gründen ihre Erklärungen zum Teil auf Avery et al. 1997.

gegebenenfalls durch weitere relevante Variablen ergänzt.[31] Bei der statistischen Prüfung wird mit den Prüfungsdaten ein logistisches Regressionsmodell trainiert, welches das Entscheidungsverhalten der Bank simuliert. Anschließend werden die einzelnen Variablen auf ihren statistischen Einfluss in der Regression überprüft, wobei der Minderheitenzugehörigkeit eines Antragstellers besondere Bedeutung zukommt. Ist sie statistisch signifikant, ist eine detaillierte Prüfung der Vergabepraxis der Bank geboten und auffällige Einzelentscheidungen werden unter Umständen genauer untersucht. Hier können Aspekte der Datenerhebung und -verifizierung eine Rolle spielen, da diese nicht in das Regressionsmodell integriert sind und nur durch das Modell als „verdächtig" eingestuft werden können.[32]

Dieses Fallbeispiel illustriert einige Aspekte der Integration von Statistik und lernenden Algorithmen (in diesem Fall die logistische Regression) in eine Gleichbehandlungsprüfung. Wir halten Folgendes fest: Der *Equal Credit Opportunity Act,* der *Fair Housing Act* und der *Home Mortage Disclosure Act* sind vergleichsweise „alte" Antidiskriminierungsgesetze (1974, 1968 und 1975) und müssen vor dem Hintergrund des *Civil Rights Movements* der 1960er Jahre verstanden werden. Schutzgegenstand sind deshalb insbesondere historisch benachteiligte Gruppen, auch wenn das Gesetz auch aus verfassungsrechtlichen Gründen „on its face"[33] neutral formuliert ist und *jede* Diskriminierung aufgrund von z.B. Ethnizität verbietet, nicht nur die zu Ungunsten von Minderheiten. Dies erleichtert die statistische Prüfung und erklärt, warum ähnliche Methoden nicht in gleicher Weise etwa in der Kontrolle der Zuweisung von Studienplätzen angewendet werden kann: Dort ist eine „holistische Evaluierung" unter eng umschriebenen Umständen erlaubt (wenngleich politisch umstritten),[34] die im Namen von positiver Diskriminierung die Anzahl von Studenten aus historisch benachteiligten Minderheiten erhöhen soll. Eine derartige „holistische Evaluierung" macht es sehr viel komplizierter, statistisch die korrekte Anwendung von Entscheidungskriterien zu überprüfen.

Selbst im *Equal Credit Opportunity Act* finden wir eine problematische Regelung: Anders als die anderen beiden Gesetze verbietet er auch Altersdiskriminierung. § 1691 b) (3) erlaubt die Hinzunahme von Alter als Kriterium, sofern dies a) durch ein empirisch erzeugtes Kreditscoringsystem geschieht, das beweisbar und statistisch korrekt ist, und b) älteren Bewerbern keinen „Altersmalus" zuweist. Solch ein „empirisch erzeugtes System" wird heutzutage typischerweise durch lernende Algorithmen erzeugt und ist damit der direkteste Link zwischen diesen Gesetzen und dem Thema dieser Studie. Doch verlangt das Gesetz mehr als nur „statistische Korrektheit", das Scoring muss auch *demonstrably sound* sein, d.h., die Daten, die zum Lernen verwendet werden, müssen empirisch korrekt sein. Zudem ist das Ergebnis aber auch noch durch eine harte Regel eingeschränkt: Selbst wenn es eine klare statistische Korrelation zwischen hohem Alter und Schuldnerverzug gäbe und diese etwa durch Änderungen im Risikoverhalten biologisch erklärbar wäre, darf sich das Alter nicht gegen den Antragsteller auswirken. Diese Kombination aus statistischem und regelbasiertem Scoring lässt sich nicht mehr direkt auf Diskriminierung durch ausschließlich

[31] Calem/Longhofer 2002 beschreiben in ihrem Artikel ein Fallbeispiel einer solchen Untersuchung, in der bei einer Prüfung der zweiten Stufe Daten für insgesamt 72 Variablen für jede Entscheidung gesammelt wurden. Das zur Prüfung genutzte Modell beinhaltete nach Kürzungen durch Experten schließlich nur noch 15 Variablen über die Kreditwürdigkeit der Antragsteller.
[32] Das Fallbeispiel von Calem/Longhofer 2002 hier detailliert zu reproduzieren würde den Rahmen des Gutachtens überschreiten. Wir verweisen hier abermals auf Kapitel 4.
[33] Siehe Annotation 20 - Fourteenth Amendment: Testing Facially Neutral Classifications Which Impact on Minorities [https://bit.ly/2wXAEKn].
[34] Siehe Grutter v. Bollinger, 539 U.S. 306 (2003).

statistische Modelle wie die oben angeführten testen. Im rechtsvergleichenden Teil werden wir eine Reihe ähnlicher Erfahrungen finden: Selbst innerhalb einer Rechtsordnung, und bei der Regulierung des gleichen Lebenssachverhalts, ist das Verständnis dessen, was als rechtswidrige Diskriminierung zählt, oft von Gesetz zu Gesetz unterschiedlich und nur in Teilen auf allgemeine, mathematische Aussagen reduzierbar. Insbesondere für die verschiedenen Spielarten der indirekten Diskriminierung deuten die Erfahrungen mit den Grenzen des *Equal Credit Opportunity Acts* an, dass es keine abstrakte, universale Methode der Analyse gibt. Tests, die wie hier die Entscheidungen ex post modellieren, sind stark abhängig vom Anwendungsbereich (hier der Kreditindustrie), von den spezifischen Gesetzen, die in ihm gelten, und ihrer oft durch kontingente historische und politische Faktoren beeinflussten Auslegung durch die Gerichte. Dies scheint auch auf die Gefährdungsszenarien der vorliegenden Studie anwendbar – allgemeine statistische Testverfahren können zwar häufig (aber nicht immer) einen ersten Verdacht insbesondere in Fällen direkter Diskriminierung erzeugen. Die Evaluierung, ob diese im konkreten Fall aber auch rechtswidrig ist, verlangt darüber hinaus ein Verständnis des Anwendungsbereichs, das nicht auf statistische Verfahren reduzierbar ist. Gegenüber einer allgemeinen Lösung, dem Problem diskriminierender Algorithmen gesetzgeberisch Herr zu werden, ist Skepsis geboten, zumindest insoweit versucht werden soll, die juristischen Diskriminierungsbegriffe auf rein statistische Aussagen zu reduzieren.

Wie erwähnt dienen Gesetze wie der *Equal Credit Opportunity Act* dem Schutz historisch benachteiligter und damit auch wirtschaftlich schwächerer Gruppen und müssen auch in ihrem historischen und gesellschaftlichen Kontext gesehen werden. Die hohen Kosten des Klageweges in den Vereinigten Staaten und die Methode der Kostenzuweisung (bei der im Regelfall auch der Gewinner seine Kosten tragen muss) machten eine Durchsetzung ausschließlich durch Einzelklagen abgewiesener Bewerber unmöglich. Obgleich Verletzungen des *Equal Credit Opportunity Acts* durch abgelehnte Bewerber sowohl durch Einzelklagen als auch durch *Class Actions* im Prinzip möglich sind, scheinen sie eine Ausnahme zu sein und vor allem nicht die oben diskutierten Evaluierungsmethoden zu benutzen. Sehr viel häufiger sind Überweisungen zum Department of Justice, welches dann in Klageverfahren Schadensersatz und Strafschadensersatz verlangen kann.[35] Im Jahr 2016 wurden etwa 18 „*Fair Lending*"-Ermittlungen durchgeführt, von denen 7 zu Klagen und 6 zu außergerichtlichen Vergleichen im Werte von 37 Millionen US-$ führten.[36] *Fair Lending Examinations* spielen damit eine doppelte Rolle: Werden sie nicht bestanden, sind sie ein starker, aber widerlegbarer Beweis, dass ein Gesetzesverstoß vorlag (in den 7 Verfahren argumentierten die Banken, dass die Entscheidungen letztendlich begründet waren). Eine bestandene Prüfung hingegen schützt als widerlegbare Vermutung insbesondere vor Class Actions, aber auch Einzelklagen, was die geringe Anzahl erfolgreicher Klagen miterklärt. Im Rahmen der in Kapitel 6.2 entwickelten rechtsvergleichenden Systematik verbinden sie daher das Aufsichtsmodell mit dem der Beweislastzuweisung.

3.4.5 Algorithmen als Mittel zum Aufdecken von Diskriminierung

Die oben angeführten Strukturen wurden historisch entwickelt, um menschliche Entscheidungen und regelbasierte Expertensysteme zu kontrollieren, nicht lernende Algorithmen. Lernende Algorithmen dienen hier also der Kontrolle der Entscheidungsfindung und sind „agnostisch" hinsichtlich der Methodik, die Banken und Kreditinstitute zur

[35] Cook 1997.
[36] The Attorney General's 2016 Annual Report to Congress Pursuant to the Equal Credit Opportunity Act Amendments of 1976, zuletzt besucht am 26.07.2018 [https://bit.ly/2Qg5kyx].

Entscheidungsfindung benutzen, durch lernende Algorithmen, regelbasierte Scoringsysteme oder manuell. Lernende Algorithmen stellen daher nicht notwendigerweise (nur) eine Herausforderung für den Verbraucherschutz dar, sie ermöglichen auch eine effizientere Ermittlung durch Prüfungsbehörden sowie Dritte. In einer großen Anzahl der Fälle, in denen diskriminierende Algorithmen zur Kenntnis der Öffentlichkeit gebracht wurden, geschah dies durch eine Analyse von Wissenschaftlern, Journalisten oder NGOs, die ähnliche Methoden zur Rekonstruktion von Entscheidungen auf Grundlage der berichteten Datensätze benutzen wie die *Fair Lending Assessments*.

So entdeckte etwa ProPublica, eine dem kritischen Journalismus verpflichtete Stiftung, die inhärente Diskriminierung in dem von vielen US-Staaten zur Risikoeinstufung bei Strafzumessung und Bewährungsentscheidungen benutzten COMPAS-System durch die Analyse von 7000 Risk Scores in Broward County, Florida.[37] Für Kreditscoring war es eine Studie einer ähnlichen Stiftung, Reveal, die über ein Jahr über 31 Millionen Entscheidungen, mit einer Kombination aus Methoden, die von staatlichen Stellen in der *Fair Lending Analysis* verwendet werden, und zusätzlichen fortgeschrittenen statistischen Algorithmen analysierte. Das Ergebnis war die Feststellung einer anhaltenden Ungleichbehandlung ethnischer Minderheiten.[38] Es ist der Erfolg derartiger Ermittlungen, der die Frage aufwirft, wie weit es unter Schutz legitimer Wirtschaftsinteressen und der Privatheit der Antragsteller möglich ist, die relevanten Datenmengen weiteren Kreisen zugänglich zu machen.

So wird etwa auch in Großbritannien der Gedanke einer Algorithmentreuhand diskutiert, wie wir unten sehen werden. Offenlegungspflichten werden aber eine zentrale Rolle in der Algorithmenregulierung spielen, da erst diese die statistische Analyse ermöglichen und so „Big Data mit Big Data bekämpft“ werden kann. So konnte die Reveal-Studie nicht direkt Score-Daten verwenden, da diese nicht öffentlich gemacht werden. Die American Bankers Association hat sich gegen weitere Offenlegungspflichten ausgesprochen und zitiert außer den Kosten für ihre Mitglieder auch Sorgen über die Datensicherheit und den Schutz der Privatheit ihrer Kunden.[39] Gerade die Reveal-Studie zeigt, wie sehr viel schneller die Wissenschaft neue Analysemethoden entwickelt, als diese von offizieller Seite aufgenommen werden können, wobei sich aber auch die Frage der „Prüfer der Prüfer“ stellt. Wenn wie in unserem Beispiel Maschinelles Lernen zur Überprüfung von Entscheidungsgerechtigkeit verwendet werden soll, müssen diese Methoden selber wieder evaluiert, getestet und gegebenenfalls akkreditiert werden.

Die Reveal-Studie bestätigt eine Reihe ähnlicher Analysen: *Fair Lending Examinations* waren nur bedingt erfolgreich, den Willen des Gesetzgebers zu einer gerechteren Kreditvergabe umzusetzen.[40] Es ist aber schwer zu beurteilen, ob sich dies durch den verstärkten Einsatz von lernenden Algorithmen verbessert oder verschlimmert hat. Der Bankensektor zumindest nimmt die geringe Anzahl erfolgreicher Klagen als Zeichen dafür, dass es keine systematischen Probleme gibt (und nicht die Rechtsdurchsetzung

[37] Website ProPublica: Machine Bias - There’s software used across the country to predict future criminals. And it’s biased against blacks, 23.05.2016, zuletzt besucht am 26.07.2018 [https://bit.ly/1XMKh5R].

[38] Glant /Martinez: For people of color banks are shutting the door to homeownership [https://bit.ly/2NtWTRD].

[39] American Bankers Association, 2017; Morgan Chase machte eine ähnliche Einreichung und verlangte zudem, dass Daten auch nicht für wissenschaftliche Studien freigegeben werden sollen, siehe JP Morgan Chase *Comment Letter on Proposed Amendments to Regulation C RIN 3170 AA10* [https://bit.ly/2NxJLen].

[40] Siehe etwa mit weiteren Nachweisen Rice/Swesnik, 2013.

unzureichend ist).[41] Eine zugegebenermaßen kursorische Analyse der Literatur findet dafür keine Anzeichen. Neue Probleme durch den Einsatz von Algorithmen schafft das Gesetz aber zumindest mit einer seiner Vorschriften. Abgelehnte Bewerber haben ein Auskunftsrecht auf Begründung der Entscheidung. Wie auch die ähnliche Diskussion zum „Recht auf Erklärung“ unter der Datenschutz-Grundverordnung stellt sich hier die Frage, wie dieses Recht praktisch umgesetzt werden kann, wenn Maschinelles Lernen mit großen Datensätzen zu opaken Entscheidungen führt. Dies könnte zu einer interessanten vergleichenden Analyse des Erfolgs eines „Rechts zur Erklärung“ im Verbraucherschutzrechts führen, im Unterschied zu dem hier bevorzugten „Recht auf Analyse“[42], doch scheint es in den USA weder Fallrecht noch akademische Analyse zu diesem Anspruch unter dem *Fair Lending Act* und seiner Zukunft im algorithmischen Entscheiden zu geben.

[41] So etwa das ABA-Statement hier [https://bit.ly/2CDBKQW].
[42] Siehe dazu, unter dem Gesichtspunkt der Feststellung fehlerhafter Beurteilungen, Kapitel 7.2.

4 Algorithmische Entscheidungen aus technischer Sicht

4.1 Einführung in Maschinelles Lernen und ADM

Maschinelles Lernen (ML), oder *Machine Learning*, ist ein Teilgebiet der Forschung im Bereich der Künstlichen Intelligenz mit starkem Bezug zu angewandter Statistik und mathematischer Optimierung. Es existieren verschiedene Definitionen von Maschinellem Lernen. Nach der gebräuchlichsten bezeichnet ML die Forschung und Anwendung von Algorithmen, die eine bestimmte Aufgabe bewältigen und ihre Leistung/Performanz durch eine Form von Erfahrung verbessern.[43] Diese Erfahrung wird typischerweise durch sogenannte Trainingsdaten bereitgestellt, also eine Menge von Problem-/Aufgabeinstanzen mit „korrekten Lösungen", von denen der Algorithmus lernen kann. In solchen Fällen spricht man auch vom sogenannten *Supervised Machine Learning*.[44]

Die Aufgabe von Interesse ist in den meisten Fällen die Vorhersage des Werts einer Zielvariablen (statistische Terminologie: „abhängige Variable") von einer Menge Eingabevariablen (statistische Terminologie: „unabhängige Variablen"). Anhand der Natur der Zielvariablen unterscheiden sich verschiedene Arten von ML. Ist das Ziel der Vorhersage kategorisch (z.B. eine Ja-Nein-Entscheidung zur Kreditvergabe oder eine Einstufung des Verbrauchers in eine von mehreren möglichen Risikogruppen), spricht man von *Klassifikation*. Ist die Zielvariable hingegen ein quantitativer Zahlenwert (z.B. die automatische Ermittlung eines Preises in einem Onlineshop anhand von Kundendaten), handelt es sich um eine *Regression*.

In den hier betrachteten *Supervised Machine Learning*-Methoden trainiert also ein bestimmter Lernalgorithmus anhand von Trainingsdaten ein Modell, welches dann für strukturell gleiche Daten Zielwerte vorhersagen kann. Vereinfacht ausgedrückt besteht ein Modell hierbei aus einer Menge im Trainingsprozess kalibrierter Parameter, die von einem modellspezifischen Vorhersagealgorithmus (bzw. einer mathematischen Gleichung) mit den Eingabedaten kombiniert werden, wodurch der neue Zielwert ermittelt wird. Verschiedene ML-Modelle haben hierbei verschiedene Eigenschaften wie Modellierungskapazität (d.h. die Fähigkeit, komplexe Muster in den Trainingsdaten zu erkennen), Anforderungen an Ressourcen zum Training (Rechnerzeit und Arbeitsspeicher), Menge benötigter Trainingsdaten, Annahmen über die Struktur der Datenrepräsentation, Sensibilität gegenüber bestimmten Phänomenen in der Verteilung der Daten etc. Von besonderem Interesse ist hierbei das Ausmaß der Möglichkeit, das gelernte Modell (also die Parameter in Kombination mit dem Vorhersagealgorithmus) manuell zu inspizieren und den Einfluss der einzelnen Eingabevariablen auf die Vorhersage quantitativ und/oder qualitativ zu interpretieren. Eignet sich ein Modell gut für eine solche manuelle Untersuchung, spricht man typischerweise von einem *Whitebox*-Modell. Wenn jedoch das Modell lediglich unter erheblichem Zeitaufwand und/oder nur mit der Hilfe von ML-Expertise interpretiert werden kann, wird es als *Blackbox*-Modell bezeichnet. Zwischen diesen beiden Polen existiert

[43] Mitchell 1997, S. 2.
[44] Weitere Teilgebiete sind Mustererkennung ohne Trainingsdaten (*Unsupervised Machine Learning*) und diverse Mischformen (*Semi-Supervised Machine Learning*). Eine umfassende Darstellung ginge weit über den Umfang der Studie hinaus.

folglich eine Bandbreite an leicht und schwer interpretierbaren Modellen mit zahlreichen mehr oder weniger zugänglichen Methoden der Sichtbarmachung.

Es existiert eine Vielzahl verschiedener Algorithmen für Maschinelles Lernen.[45] Im Bereich Kreditscoring werden in einer Publikation von 2015 über 40 verschiedene Algorithmen und Verfahren für Maschinelles Lernen unterschieden.[46] Andererseits haben wir in unseren Gesprächen erfahren, dass die meisten Unternehmen beim Kreditscoring im Wesentlichen die gleichen, bewährten Verfahren verwenden (z.B. logistische Regression). Im Kontext dieser Studie gehen wir daher nur zu Einführungszwecken auf ausgewählte Modelle ein. Das Ziel der Ausführungen ist, (1) dem Leser Grundlagen in der Repräsentation von Daten als vieldimensionale Vektorräume zu vermitteln, (2) das Konzept von Modellparametern im Spannungsfeld zwischen *Whitebox*- und *Blackbox*-Modellen zu veranschaulichen sowie (3) einen Kurzabriss über Methoden zu geben, mit denen Modelle, bzw. ihre Parameter, automatisch von Daten „gelernt" werden können. Aus Letzterem lassen sich bestimmte Lernverhalten ableiten, die Algorithmen/Modelle zu „diskriminierendem" Verhalten verleiten.[47]

4.1.1 Grundlagen linearer und logistischer Regression

Die wohl einfachste Form selbstlernender, parametrisierter Modelle sind in der Statistik schon seit langer Zeit gebräuchliche Regressionsmodelle. In der Grundform handelt es sich dabei um eine Zielvariable y, deren Wert von den Eingabevariablen $x_1, \ldots, x_n$ berechnet wird, indem jede Eingabevariable mit einem Gewichtsparameter $w_1, \ldots, w_n$ multipliziert und schließlich alles aufsummiert wird. Dies ergibt eine sogenannte „lineare Regression":

$$\hat{y} = w_1 x_1 + w_2 x_2 + \ldots + w_n x_n$$

Mit Hilfe einer Regression lassen sich beispielsweise Mietpreise von Wohnungen anhand der Quadratmeterzahl, der Anzahl der Zimmer und Bäder sowie der Entfernung zur Stadtmitte berechnen. Das Modell lernt, indem es solche Parameter ermittelt, die die Gesamtabweichung (*Error* oder *Loss*) des durch die Gleichung ermittelten Preises minimieren. Es existieren verschiedene Methoden zum Finden dieses Optimums im Raum aller möglichen Parameterkombinationen mit verschiedenen Eigenschaften. Eine Einführung in solche Algorithmen würde den Rahmen dieser Studie übersteigen. Nach der Optimierung lassen sich die Parameter des Modells untersuchen und gegebenenfalls sachliche Schlüsse daraus ziehen. Zum Beispiel würde die Anzahl der Zimmer mit einem Gewichtsparameter versehen, der bei der Untersuchung darüber Aufschluss gibt, mit welchem Faktor zusätzliche Zimmer zum Gesamtwert der Wohnung beitragen. Hierdurch ist die lineare Regression ein Beispiel eines einfachen und *erklärbaren* Modells statistischen Lernens. Das Maß der Erklärbarkeit ist semantisch mit den Eingabevariablen verknüpft, da wir die Parameter als quantitativen Einfluss der jeweiligen Variablen auf die Gesamtvorhersage verstehen können.

Die lineare Regression ist eine übliche Technik zum Vorhersagen von Zahlenwerten. Wenn stattdessen ein binärer Wert (entweder 0 oder 1) gesucht wird (und damit ein

[45] Füser 2013.
[46] Lessmann et al. 2015.
[47] An dieser Stelle sei angemerkt, dass im technischen Teil dieses Gutachtens terminologisch nicht von einer „Diskriminierung" gesprochen werden kann, sondern lediglich von einer Ungleichbehandlung von verschiedenen Gruppen innerhalb von Datensätzen.

Klassifikationsproblem vorliegt), lässt sich die Formel entsprechend durch eine sogenannte „logistische" Funktion adaptieren:

$$\hat{y} = \sigma(w_1 x_1 + w_2 x_2 + ... + w_n x_n)$$

Hier wird eine lineare Regression in eine nichtlineare Funktion σ eingebettet, die Zahlenwerte beliebiger Größe in ein begrenztes Intervall (typischerweise zwischen 0 und 1) projiziert. Liegt diese über- oder unterhalb eines Schwellenwerts, wird entweder eine positive oder negative Vorhersage getroffen und entsprechend eine binäre Entscheidung modelliert (z.B. zur Kreditwürdigkeit einer Person). Dieses Modell wird allgemein als „logistische Regression" bezeichnet.[48] Es existieren verschiedene logistische Funktionen, deren genaue Funktionsweise hier nicht dargelegt werden muss. Wichtig ist, dass auch hier Gewichtsparameter einer linearen Gleichung gelernt werden, indem man die Differenz des vorhergesagten Werts mit dem richtigen Trainingswert minimiert (z.B. eine Einschätzung der Kreditwürdigkeit zwischen 0 als nicht kreditwürdig und 1 als uneingeschränkt kreditwürdig). Auch in einer logistischen Regression können nach dem Training die gelernten Parameter untersucht und anhand ihrer zugeordneten Eingabevariablen helfen, das Verhalten des Modells zu erklären.

4.1.2 Logistische Regression in der Praxis

Lineare und logistische Regression sind klassische Modelle aus der Statistik und finden seit sehr langer Zeit praktische Anwendung, insbesondere bevor ML eine gesteigerte Aufmerksamkeit zukam.[49] Besondere Bedeutung kommt hierbei der Möglichkeit zu, die gelernten Koeffizienten zu untersuchen, die das Modell transparent und zu einem mächtigen Werkzeug zur Datenanalyse machen. Es wird seit geraumer Zeit in den amerikanischen Gerichten als Beweismittel für ethnische Ungleichbehandlung verwendet.[50]

Interviews: Aus der Erfahrung unserer Gespräche mit Anwendern von ADM-Technologie haben wir Grund zu der Annahme, dass die logistische Regression im Bereich von Verbraucherscoring das in der Praxis am häufigsten verwendete statistische Modell ist. Als Gründe wurden uns zwei wesentliche Aspekte genannt. Zum einen produziert sie, wie oben dargestellt, interpretierbare Parameter. Zum anderen ist die Vorhersagegenauigkeit im Vergleich mit komplexeren Modellen sehr gut. In einem Gespräch wurde uns erklärt, komplexere (und damit schwieriger nachvollziehbare) Modelle basierend auf den vorliegenden Verbraucherdaten würden nur eine marginal höhere Genauigkeit erzielen, womit das einfachere und transparente Modell vorzuziehen sei.

[48] Klarstellung: Wie oben erwähnt bezeichnet „Regression" ein Modell mit einer reellen Zahl als Zielvariable und „Klassifikation" eine kategorische Zielvariable. Die logistische Regression modelliert die Wahrscheinlichkeit, dass die Dateninstanz einer bestimmten Kategorie angehört. Wir behandeln sie daher als äquivalent zu einem Klassifikationsmodell. Für eine Umfangreiche Darstellung der logistischen Regression siehe Hastie et al. 2009, Kapitel 4.4.

[49] Manche Quellen grenzen daher die logistische Regression als statistisches Modell von neueren *Machine-Learning*-Modellen ab. Wir widersprechen dieser Auffassung und halten die Regression mit der unserem Dafürhalten nach herrschenden Ansicht für einen Grundbaustein von ML und komplexeren Modellen. Sie wird auch typischerweise in ML-Kursen als Grundlage gelehrt.

[50] Vgl. als frühe Übersicht Finkelstein 1980.

Lineare und logistische Regression sind auf sogenannte *lineare* Modelle beschränkt.[51] Dies bedeutet, dass der Lernalgorithmus jede Eingabevariable nur einzeln und nicht in Kombination berücksichtigen kann. Beispielsweise könnte die Tatsache, dass ein Kreditbewerber Kinder hat, generell ein negatives Signal sein (wegen höherer Lebenshaltungskosten). Wenn jedoch die Kinder bereits erwerbstätig sind, könnte sich dies positiv auf die Kreditwürdigkeit auswirken (da keine Unterhaltskosten mehr anfallen oder sogar eine höhere Solvenz der Familie vorliegt). Eine solche *nichtlineare* Interaktion zwischen dem Vorhandensein und der Erwerbstätigkeit der Kinder lässt sich in einer linearen oder logistischen Regression nicht ohne weiteres modellieren.[52]

4.1.3 Komplexere ML-Modelle

Im Bereich Maschinellen Lernens existiert eine Vielzahl von Modellen und Lernalgorithmen mit verschiedenen Funktionsweisen, Eigenschaften und typischen Anwendungsgebieten. Wir geben hier einen kurzen Überblick einiger Modelle für das im Gutachten relevante *Supervised Learning* (also Lernen anhand von korrekten Trainingsbeispielen). Ziel der Ausführungen ist hierbei nicht eine technische Spezifikation, sondern die Illustration der jeweiligen Funktionsweise und Interpretationsmöglichkeiten im Hinblick auf die inhaltliche Überprüfung.

Das zugrundeliegende Paradigma von *Naive Bayes*[53] ist es, Zusammenhänge zwischen Attributen in Form von Wahrscheinlichkeiten darzustellen. *Naive-Bayes*-Verfahren beruhen auf der Annahme, dass Attribute voneinander unabhängig sind. Dies ist eine starke Annahme und nur selten richtig, jedoch funktioniert dieses Verfahren in der Praxis sehr gut und kann sehr effizient berechnet werden. Die gelernten Parameter der Einflussnahme über Wahrscheinlichkeiten können von Menschen gelesen und analysiert werden. Komplexere grafisch darstellbare Zusammenhänge von Wahrscheinlichkeiten können mit Hilfe sogenannter *Bayesian Networks*[54] modelliert werden. Dort bilden die Parameter ebenfalls bedingte Wahrscheinlichkeiten, jedoch ist die Netzstruktur auch von Daten lernbar, was die Erklärbarkeit unter Umständen erschwert. *Support Vector Machines*[55] transformieren Attribute in einen hochdimensionalen Vektorraum und errechnen dort sogenannte „Stützvektoren“ einer Entscheidungsgrenze, anhand derer unterschiedliche Klassen voneinander getrennt werden können. Ihre gelernten Parameter sind aufwändig zu interpretieren, speziell wenn sogenannte Kernel-Projektionen verwendet werden. Entscheidungsbäume[56] (*Decision Trees*) partitionieren den Raum in möglichst homogene Bereiche durch eine Verzweigung auf der Basis von Attribut-Wertprüfungen. Hier besteht das gelernte Modell nicht aus einer Menge von Parametern, sondern aus der Baumstruktur selbst. Einzelne Bäume sind typischerweise leicht interpretierbar. *Random Forests* hingegen kombinieren mehrere Entscheidungsbäume in ein übergeordnetes Modell, was die Interpretierbarkeit unter Umständen stark einschränkt.

---

[51] Die logistische Regression verwendet zwar eine nichtlineare logistische Funktion, jedoch ist die durch das Modell gegebene Entscheidungsgrenze zwischen den beiden Klassen.

[52] Lineare Modelle können durch Kernel-Methoden auf nichtlineare Interaktionen erweitert werden. Dies geht oft mit einer Vervielfachung der zu lernenden Parameter einher, wodurch die Interpretierbarkeit des Modells unter Umständen erschwert werden kann. Der Einsatz solcher Methoden kam in den im Rahmen der Studie geführten Interviews nicht zur Sprache.

[53] Siehe Kapitel 13 und 14 in Russell/Norvig 2010.

[54] Siehe Kapitel 14 in Russell/Norvig 2010.

[55] Siehe Kapitel 18.9 in Russell/Norvig 2010.

[56] Siehe Kapitel 18.3 in Russell/Norvig 2010.

Neuronale Netzwerke[57] sind komplexe Gleichungssysteme, die in ihrer Struktur an die Funktionsweise von Neuronen im menschlichen Gehirn angelehnt sind. Stark vereinfacht ausgedrückt handelt es sich dabei um eine große Menge ineinander verschachtelter logistischer Regressionsgleichungen, die mit zunehmender Größe eine sehr hohe Lernkapazität haben. Besondere Bedeutung kommt hierbei den sogenannten *hidden nodes* zu, die nichtlineare Interaktionen lernen können. Neuronale Netzwerke sind seit vielen Jahrzehnten bekannt, fanden aber erst in den vergangenen zehn Jahren breite Anwendung.[58] Die produktivsten Anwendungsfelder sind Bild- und Textverarbeitung, jedoch können sie in jedem statistischen Lernverfahren eingesetzt werden. Neuronale Netze sind gegenwärtig in der Forschung und den Medien sehr populär und Softwarepakete, die deren Anwendung erleichtern, werden immer zugänglicher. Trainierte neuronale Modelle sind je nach Komplexität und Struktur nur schwer erklärbar, da sie über sehr große Mengen Parameter verfügen. Die Erklärbarkeit solcher Modelle ist ein Feld aktueller Forschung.

4.1.4 Praktische Grenzen der Erklärbarkeit

Diese oberflächlichen und stark vereinfachenden Erklärungen legen bereits nahe, dass eine *One-size-fits-all*-Analyse verschiedener ML-Modelle nicht möglich ist. Verschiedene Lernalgorithmen produzieren durch Training verschiedenartige Voraussagemodelle, die jeweils durch entsprechende Methoden mehr oder weniger für den Menschen verständlich nachvollzogen werden können.

Wie weiter unten (insbesondere Kapitel 4.4) gezeigt wird, existiert eine Bandbreite an Möglichkeiten, um das Verhalten von ML-Modellen, die ADM-Systemen zugrunde liegen, zu erklären bzw. erklärbar zu machen. Es gibt jedoch auch technische Limitationen, die es eben nicht erlauben jedes ML-Modell gleichermaßen zu erklären. Dies liegt schlicht und einfach an der Verschiedenheit der jeweiligen Technologie.

Ein überwiegender Teil bestehender Literatur im Bereich *„Fairness in Machine Learning“* behandelt die „Fairness“ von Klassifikationsmodellen zu binären Ja-Nein-Entscheidungen (z.B. Entscheidungen über Kreditwürdigkeit), wenn die Personen bestimmten Gruppen angehören und die Entscheidungen unter Gleichbehandlung der Gruppen getroffen werden sollen. Aufgrund der Uneinheitlichkeiten geht ein wesentlicher Teil dieser Forschung von einem intransparenten *Blackbox*-Modell aus und arbeitet mit verschiedenen Techniken, um Fairness zu analysieren und auf verschiedene Art und Weise zu fördern.

4.2 Entstehung von Ungleichbehandlung durch lernende Algorithmen

4.2.1 Von unausgewogenen Daten zur unausgewogenen Vorhersage

Das Risiko der Ungleichbehandlung ergibt sich bei der automatisierten Entscheidungsfindung durch das Training des Klassifikationsmodells aufgrund zweier Faktoren: (1) der unmittelbaren oder mittelbaren Information über die Gruppenzugehörigkeit

[57] Siehe Kapitel 18.7 in Russell/Norvig, 2010.
[58] Dies folgte aus der Entdeckung, dass neuronale Netzwerke effizient mittels spezieller Grafikhardware trainiert werden können, vgl. Rajat et al. 2009.

im Datensatz sowie (2) einer statistischen Unausgewogenheit in der Verteilung des Werts der Zielvariablen über die Personen in den entsprechenden Gruppen. Beispielsweise könnten in einem Datensatz über Kreditwürdigkeit drei Viertel aller als kreditwürdig erachteten Personen weiblich sein und nur ein Viertel männlich.

In diesem Fall könnte ein Lernalgorithmus statistisch schließen, dass die Wahrscheinlichkeit der Kreditwürdigkeit vom Geschlecht der antragstellenden Person abhängt. Dies kann unter Umständen dazu führen, dass der Algorithmus dieses Ungleichgewicht verstärkt und nun vier Fünftel der vom Modell als kreditwürdig eingestuften Personen weiblich sind und nur ein Fünftel männlich. Dieses Problem einer unzulässigen oder unausgewogenen Verwendung der Information zur Gruppenzugehörigkeit sowie der Verstärkung von Unausgewogenheit in den Trainingsdaten (*„Bias Amplification"*[59]) im Laufe des Trainingsprozesses ist ein technisches Kernproblem der Forschung im Bereich der Diskriminierung durch Algorithmen sowie ihrer Prüfung und Vermeidung.

Hieraus folgt ein grundlegendes Dilemma im Bereich sogenannter *Data-Driven Systems*, also Systemen, die für einen wesentlichen Teil ihrer Funktionalität statt mit manuell programmiertem Expertenwissen mit statistischen Inferenzen auf einem Referenzdatensatz arbeiten. Wenn die zugrundeliegenden Daten inhärente Unausgewogenheiten zwischen repräsentierten Gruppen oder Konzepten haben, wird das statistische Modell diese lernen, wenn nicht im Lernprozess entsprechende Vorkehrungen getroffen wurden. Die am meisten diskutierten Beispiele betreffen bekanntermaßen ADM-Systeme für rechtlich unmittelbar bedeutsame Entscheidungen über Individuen, wie beispielsweise automatisierte Empfehlungen zu Bewährungsentscheidungen, die Gefangene mit bestimmtem ethnischem Hintergrund als prinzipiell risikobehafteter einstufen[60]. Der Vollständigkeit halber sei an dieser Stelle angemerkt, dass das Phänomen im Zusammenhang mit sozialen Gleichstellungsimperativen durchaus subtil sein kann. Beispielsweise wurde erkannt, dass in einem in der Forschung oft benutzten Bildbeschreibungsdatensatz rund 47 % aller Verben jeweils mit Bildermengen assoziiert wurden, in denen eine Unausgewogenheit in Bezug auf das Geschlecht der abgebildeten Personen vorlag. So war auf zwei Dritteln aller Bilder mit dem Beschreibungsverb „Kochen" eine Frau zu sehen und auf einem Drittel ein Mann.[61]

4.2.2 Vermeiden von Ungleichbehandlung im Modell-Trainingsprozess

In der Forschungsliteratur existieren Ansätze[62], den Einfluss solcher Unausgewogenheiten zum Zeitpunkt des Modelltrainings zu begrenzen oder zu eliminieren, indem Ausgewogenheit als eine bestimmte Verteilung in den Trainingsdaten definiert wird. Der Trainingsdatensatz kann dann entsprechend ausbalanciert werden, indem man entweder kontrolliert Datenpunkte der Mehrheitsgruppe nicht berücksichtigt (*Undersampling*) oder durch bestimmte Prozeduren künstlich zusätzliche Datenpunkte der unterrepräsentierten Gruppe generiert (*Oversampling*). Beide Techniken gehören zwar zum allgemeinen Repertoire in der Arbeit von ML-Modellen, haben jedoch auch Nachteile (z.B. unter Umständen eine reduzierte Genauigkeit des trainierten Modells) und führen nicht in allen Situationen zum gewünschten Ergebnis. Unserer Ansicht nach eignet sich die Benutzung solcher Verfahren mithin zum gegenwärtigen Zeitpunkt nicht als unqualifizierte allgemeine Anforderung an die Entwickler und Verwender von ADM-Systemen im Sinne einer Regelung

[59] Zhao et al. 2017.
[60] Siehe Angwin et al. 2016.
[61] Zhao et al. 2017.
[62] Z.B. Feldmann et al. 2015; Zhao et al. 2017.

oder Zertifizierung. Sie gehören jedoch zum methodischen Standardrepertoire bei der Erstellung statistischer Modelle und sind bei der Gestaltung von Prüfungs- bzw. Auditverfahren für ADM definitiv zu berücksichtigen.

4.2.3 Direkte und indirekte Einflussnahme

Wenn ein geschütztes Attribut, also ein solches, das bei seiner Berücksichtigung zu diskriminierenden Effekten führen kann (z.B. das Geschlecht des Antragstellers), mit der Zielvariablen korreliert, wird ein ML-Modell es hochwahrscheinlich zur Vorhersage aufgreifen. Wie bereits dargelegt, kann es hier zu Ungleichbehandlungen im trainierten Modell kommen. Das klassische Beispiel aus der amerikanischen Literatur im Bereich automatisierter Bewährungsentscheidungen ist, dass die Rassenzugehörigkeit statistisch mit der Rückfallquote in Beziehung steht. Die einfachste Sicherheitsvorkehrung ist die Nichtberücksichtigung des geschützten Attributes („*Fairness by Blindness*“). Dies ist intuitiv und erscheint zunächst effektiv, jedoch ergeben sich folgende zwei Probleme.

Zum einen gibt es andere, harmlos erscheinende Attribute, die das geschützte Attribut ganz oder teilweise beinhalten („Proxyvariable“), so dass die Ungleichbehandlung mittelbar trotzdem vom Modell gelernt wird. Beispielsweise korreliert ethnischer Hintergrund oft stark mit Wohnort und Postleitzahl. Information zur Gruppenzugehörigkeit ist oftmals über mehrere Attribute verstreut. Zum anderen wird ein statistisches Modell durch das Entfernen von mit dem Zielwert korrelierenden Variablen hochwahrscheinlich immer ungenauer in seinen Voraussagen. Je mehr Attribute aufgrund ihres Zusammenhangs mit der Gruppenzugehörigkeit entfernt werden, umso mehr Genauigkeit wird verloren. In der *Fair Machine Learning*-Literatur wird die Einbuße von Genauigkeit zur Einhaltung von Fairnesskriterien typischerweise als „*pseudo-regret*“ oder einfach „*regret*“ bezeichnet.[63]

Praxis: Wir führten ein Gespräch mit der Schufa Holding AG, die Verbraucherdaten sammelt und durch ein Regressionsmodell erstellte Kreditwürdigkeitsbewertungen verkauft. Uns wurde erklärt, dass Attribute, deren Verwendung durch das deutsche Recht direkt verboten wird (z.B. geschlechtliche Orientierung, ethnische Herkunft etc., vgl. Art. 9 Abs. 1 DS-GVO) selbstverständlich nicht erhoben werden. Ebenso wenig würden Attribute (z.B. Marketingdaten („welche Zeitschriften werden abonniert?), über die auf verbotene Attribute geschlossen werden könnte), weder erhoben noch beim Modelltraining verwendet werden. Verwendung fänden ausschließlich Daten, die einen statistisch eindeutigen Bonitätsbezug aufweisen, wie z.B. Kreditnutzung oder Zahlungsstörungen, sodass eine ungerechtfertigte Unterscheidung ausgeschlossen sei.

Der Einfluss von einzelnen Attributen auf die Vorhersagegenauigkeit eines Modells lässt sich z.B. durch eine zufällige Perturbierung feststellen. Hierbei werden Daten durch das Modell vorhergesagt und wird die Genauigkeit quantitativ festgestellt. Danach wird in den gleichen Daten das Attribut von Interesse in jedem Datensatz zufällig verändert. Für diesen perturbierten Datensatz werden nun abermals die Zielwerte durch das Modell vorausgesagt und wird die Genauigkeit quantifiziert. Die Differenz in der Genauigkeit zwischen den Originaldaten und dem Datensatz mit dem zufällig veränderten Attribut bildet dann ein quantitatives Maß für den Einfluss der Variablen auf das Voraussageverhalten.

Nehmen wir beispielsweise an, dass wir ein trainiertes Modell zur Bestimmung der Kreditwürdigkeit eines Antragstellers hinsichtlich des Einflusses des Geschlechts

[63] Siehe Definitionen in beispielsweise: Joseph et al. 2016.

untersuchen wollen. Wir perturbieren die Daten, indem wir in jedem Datensatz das Geschlecht des Antragstellers mit einer Wahrscheinlichkeit von 0,5 ändern (von männlich zu weiblich bzw. umgekehrt). Ist das Geschlecht für die Voraussage relevant, wird sich das Modell mit diesen veränderten Daten anders verhalten und weniger genau sein. Siehe dazu auch das Beispiel in Kapitel 4.4.3.5. Die Differenz der Genauigkeit im Vergleich zu der Vorhersage des Modells mit den Originaldaten ist eine mögliche Quantifizierung des Einflusses des Geschlechtsattributs auf das Modell. Es sei angemerkt, dass dieses Verfahren fehleranfällig ist (beispielsweise im Hinblick auf bereits erklärte redundante Information durch Proxyvariablen) und in der Literatur Alternativen/Varianten diskutiert werden.[64]

4.2.4 Zwischenfazit

Aus den bisherigen Ausführungen in diesem Kapitel ergibt sich folgendes Zwischenfazit:

- *Machine Learning*- bzw. (statistische) ADM-Modelle können lernen, ungleiche Entscheidungen für gleich zu behandelnde Gruppen zu treffen, wenn in den Trainingsdaten die Gruppenzugehörigkeit mit der Zielvariablen korreliert.
- Wird die Gruppenzugehörigkeit im Modell nicht verwendet, kann dies entsprechend negative Auswirkungen auf die Genauigkeit der Vorhersage haben.
- Die Gruppenzugehörigkeit kann auch implizit durch andere sogenannte Proxyvariablen repräsentiert sein, die gegebenenfalls dann ebenfalls ausgeschlossen werden müssen.
- Oft existiert ein Zielkonflikt zwischen dem Anspruch der inhaltlichen Richtigkeit der algorithmischen Entscheidungen (Gefährdungsszenario 1) und der Vermeidung von Ungleichbehandlung durch solche (Gefährdungsszenario 2), da sich die Nichtverwendung von korrelierenden Variablen negativ auf die Genauigkeit auswirkt.
- Der Einfluss bestimmter Variablen auf die Vorhersage eines Modells lässt sich durch Veränderung der Variablen in einem Datensatz und Berechnen des Ergebnisses ermitteln.

4.3 „Fairness“ im Maschinellen Lernen

Das Forschungsgebiet *Fair Machine Learning* beschäftigt sich seit einigen Jahren mit der Frage, wie die Technologie um AI, Big Data und ML auf verantwortliche, transparente, sozialverträgliche Weise benutzt werden kann und sollte.[65] Ein Teil der Arbeit in diesem Bereich gilt der Sichtbarmachung, Messung und Vermeidung von Ungleichbehandlung durch ML-Modelle. Trotz der Neuheit dieses Gebiets sind die dort bereits gewonnenen Erkenntnisse für dieses Gutachten insofern relevant, als sie Phänomene von Ungleichbehandlungen in datenbasierten Systemen formal-mathematisch klar definieren und somit die aktuelle und künftige Diskussion über und Entwicklung von datenintensiven Systemen sowie deren Regulierung fördern und strukturieren. Es sei angemerkt, dass wir den Begriff der „Fairness“ hier nicht in die juristische Fachterminologie einführen wollen,

[64] Siehe Adler et al. 2018.
[65] Die Forschungsgemeinschaft überschneidet sich mit dem Gebiet von Ethik & Künstliche Intelligenz. Forschungsbeiträge werden sowohl in den jeweiligen Fachkonferenzen und Zeitschriften als auch im Rahmen spezieller Veranstaltungen veröffentlicht. Siehe z.B. die Workshopreihe seit 2014 für *Fairness, Accountability & Transparency in Machine Learning* (FATML) [www.fatml.org/].

sondern wir ihn wie in der *Fair Machine Learning*-Literatur üblich verwenden, um die dort gewonnenen Einsichten für die juristische Diskussion nutzbar zu machen.

4.3.1 Performancemetriken für ML-Modelle

Vorab bedarf es der Erklärung von gängigen Metriken zur Beurteilung der Performanz eines ML-Modells in binärer Positiv-negativ-Entscheidung.[66] The folgende Abbildung ist eine Erweiterung einer sogenannten *Confusion Matrix* zur Bewertung eines Klassifikationsmodells.[67]

	Vorhersage positiv	Vorhersage negativ	Bedingte Gruppenfehler bzw. -genauigkeit *Conditional Procedure Error*
Tatsächlich positiv	*true positive* TP	*false negative* FN	*false negative rate* *FNR = FN / (TP + FN)*
Tatsächlich negativ	*false positive* FP	*true negative* TN	*false positive rate* *FPR = FP / (TN + FP)*
Bedingter Vorhersage-fehler bzw. -genauigkeit *Conditional Use Error*	*positive prediction error* *PPE = FP / (TP + FP)*	*negative prediction error* *NPE = FN / (FN + TN)*	*accuracy (Genauigkeit)* ACC = TN + TP / TP + FP + FN + TN *overall (procedure) error* OE = 1 - ACC

Abbildung 1: Erweiterte „Confusion Matrix“ zur Bewertung eines Klassifikationsmodells

True Positives bezeichnet die Menge an tatsächlich kreditwürdigen Personen, die auch als solche von einem Modell vorhergesagt wurden. *False Negatives* benannt die Menge der kreditwürdigen Personen, die vom System fälschlicherweise als nicht kreditwürdig eingestuft wurden. Die *False Negative Rate* ist dementsprechend die Wahrscheinlichkeit (also eine Prozentwerten äquivalente Zahl zwischen 0 und 1), mit der eine kreditwürdige Person als kreditunwürdig eingestuft wird. Die Genauigkeit (*Accuracy*) eines Klassifikationsmodells ist das Verhältnis (ebenfalls prozentgleich zwischen 0 und 1) der Anzahl korrekter Voraussagen zu der Gesamtzahl an getätigten Voraussagen.

Neben diesen Metriken kommt in der Praxis bei der Genauigkeitsbewertung von Klassifikationsmodellen im Bereich Kreditscoring unter anderem der Gini-Koeffizient zum Einsatz. Dabei handelt es sich um eine Messung der Genauigkeit des statistischen Klassifikationsmodells im Vergleich mit einem gänzlich „uninformierten“, zufälligen Entscheidungsmodell (z.B. einer 50/50-Zufallsentscheidung für/gegen die Kreditvergabe).[68]

[66] Die Terminologie für die Metriken ist im Folgenden teilweise in englischer Sprache, da nicht alle Konzepte deutsche Bezeichnungen haben bzw. diese nicht allgemein gebräuchlich sind. Bestimmte Begriffe wurden von den Autoren der Studie sinngemäß übersetzt.
[67] Im Folgenden übersetzt, paraphrasiert und gekürzt nach Berk et al. 2017.
[68] Siehe z.B. Rezac/Rezac (2011). Die Anwendung des Gini-Index kam auch in unseren Gesprächen mit Praktikern zur Sprache.

Es kann hier zunächst festgehalten werden, dass das Konzept der inhaltlichen Richtigkeit einer algorithmischen Entscheidung (SVRV-Gefährdungsszenario 1) differenziert betrachtet werden muss und das Verhalten eines Modells durch eine Verteilung über korrekte Voraussagen und verschiedene Fehlerkategorien charakterisiert wird (z.B. *false negative/positive rate*). Die Frage, ob ein Modell zwei Gruppen gleich behandelt, ist folglich nicht nur eine Frage der Nichtverwendung der Gruppenzugehörigkeit in der Entscheidung, sondern auch potenziell einer vergleichbaren Genauigkeit sowie vergleichbaren „Fehlverhaltens" in den Gruppen bei Falschprognosen.

Ein prominentes Beispiel einer Untersuchung von Fehlverhalten eines Algorithmus zwischen Gruppen ist die Untersuchung des COMPAS-Systems zur Risikoeinstufung bei Strafzumessung und Bewährungsentscheidungen in den USA durch die Nachrichtenagentur ProPublica.[69] Eine der Beobachtungen der Analyse war, dass in den verwendeten Testdaten afroamerikanische Angeklagte im Vergleich zu weißen Angeklagten mit nahezu der doppelten Wahrscheinlichkeit fälschlicherweise als hochgradig risikobehaftet eingestuft wurden, erneut eine Straftat zu begehen (Vergleich der *false positive rates*).[70]

Ausgehend von diesen Metriken erläutern wir im Folgenden eine Reihe quantitativer Fairnessbegriffe aus der Literatur und präsentieren ein Beispiel zu deren Veranschaulichung.

4.3.2 Quantitative Fairnessbegriffe für ML-Modelle

Der Schlüssel zum Verständnis von quantitativen Diskriminierungsbegriffen liegt in der Betrachtung und dem Vergleich dieser Metriken für die gleich zu behandelnden Untergruppen des Datensatzes. Berk et al. (2017) sichten vorhandene Arbeiten und unterscheiden die folgenden fünf quantitativen Gleichbehandlungsbegriffe:

- **Gleiche Genauigkeit („*Overall accuracy equality*"):** Ein Klassifikationsalgorithmus erfüllt das Kriterium der gleichen Genauigkeit, wenn er in allen geschützten (also gleich zu behandelnden) Gruppen innerhalb des Datensatzes die gleiche Vorhersagegenauigkeit hat. In diesem Fall werden Mitglieder aller Gruppen mit gleicher Wahrscheinlich richtig oder falsch eingestuft. Ein Nachteil dieses Kriteriums ist, dass keine Unterscheidung nach Art der Einstufung getroffen und daher *false positives* und *false negatives* gleich behandelt werden. Wenn also zwei gleich zu behandelnde Gruppen im Datensatz tatsächlich jeweils mehrheitlich negativ und mehrheitlich positiv bewertet werden, kann ein Algorithmus die erste Gruppe negativ und die zweite positiv diskriminieren und trotzdem in beiden Gruppen gleich genau sein.

[69] Julia Angwin, Jeff Larson, Surya Mattu und Lauren Kirchner, *Machine Bias*, ProPublica, 23. Mai 2016 [https://bit.ly/1XMKh5R]. Dieses Produkt zeigt auch eine weitere Gefahr. COMPAS war ursprünglich nur für die Unterstützung von Bewährungsentscheidungen gedacht, wird aber auch in einigen Staaten zur Strafzumessung verwendet. Diese, von den Entwicklern nicht antizipierte, Anwendung wirft zusätzliche rechtliche und faktische Probleme auf, so dass selbst wenn das Problem in der einen Anwendung zulässig sein sollte, es in der anderen eine rechtlich problematische Diskriminierung zur Folge hätte. Dies weist darauf hin, dass eine „abstrakte" Evaluierung der Korrektheit eines Algorithmus nicht immer möglich ist, sondern jede Anwendung und ihr Rechtsrahmen individuell analysiert werden müssen, selbst wenn diese wie im vorliegenden Fall sehr ähnlich sind.

[70] Jeff Larson, Surya Mattu, Lauren Kirchner and Julia Angwin, *How We Analyzed the COMPAS Recidivism Algorithm*, 23. Mai 2016 [https://bit.ly/1TGK42v].

- **Statistische Parität („*Statistical parity*"):** Statistische Parität besteht, wenn die Werte der Zielvariablen in jeder Gruppe gleich verteilt sind. In unserem Beispiel wäre dies der Fall, wenn das Mengenverhältnis von Kreditwürdigkeit und -unwürdigkeit in beiden Geschlechtern gleich wäre. Klassifikationsalgorithmen können auf unausgewogenen Daten trainiert und zur Einhaltung von statistischer Parität in der Voraussage gezwungen werden. Dies führt allerdings zu zahlreicheren *false positives* bzw. *false negatives* auf Seiten der unter- bzw. überrepräsentierten Gruppe und reduziert entsprechend die Gesamtgenauigkeit.

- **Gleiche bedingte Gruppengenauigkeit („*Conditional procedure accuracy equality*"):** Ein Algorithmus erfüllt das Kriterium der gleichen bedingten Gruppengenauigkeit, wenn er in den gleich zu behandelnden Gruppen die gleiche *false positive/negative rate* hat. In unserem Beispiel hieße dies, dass tatsächlich kreditwürdige Frauen und Männer mit der gleichen Wahrscheinlichkeit fälschlicherweise als kreditunwürdig eingestuft würden bzw. tatsächlich nicht kreditwürdige Frauen und Männer unrichtigerweise einen Kredit erhielten. Es geht also um eine Vergleichbarkeit der Fehlerwahrscheinlichkeit für *tatsächlich* kreditwürdige (bzw. nicht kreditwürdige) Personen.

- **Gleiche bedingte Vorhersagegenauigkeit („*Conditional use accuracy equality*"):** Gleiche bedingte Vorhersagegenauigkeit besteht, wenn der Algorithmus innerhalb einer bestimmten Kategorie von Voraussagen sich in allen Gruppen mit der gleichen Wahrscheinlichkeit irrt. In unserem laufenden Beispiel wäre dies erfüllt, wenn innerhalb aller als kreditwürdig eingestuften Frauen und Männern sich der Algorithmus jeweils mit der gleichen Wahrscheinlichkeit irrt (also gleicher *positive/negative prediction error*). Praktisch hieße dies beispielsweise, dass als kreditwürdig eingestufte Männer und Frauen mit der gleichen Wahrscheinlichkeit entgegen der Prognose das Darlehen nicht zurückzahlen können. In Abgrenzung zur Gruppengenauigkeit geht es um eine Vergleichbarkeit der Fehlerwahrscheinlichkeit für vom Algorithmus als kreditwürdig (bzw. nicht kreditwürdig) eingestufte Personen.

- **Gleiches Fehlerverhältnis („Treatment equality"):** Gleiches Fehlerverhältnis besteht, wenn das Verhältnis von *false positives* und *false negatives* in allen Gruppen gleich ist. Dies ist beispielsweise nicht der Fall, wenn es sich bei Fehleinstufungen von Frauen überwiegend um Kreditablehnungen handelt und bei Männern überwiegend um Kreditgenehmigungen. Ein Algorithmus sollte in seinen Irrtümern in verschiedenen Gruppen nicht unterschiedliche Fehlertendenzen haben.

Berk et al. (2017) fassen die Vereinigung aller fünf Begriffe von algorithmischer Fairness im Konzept der *total fairness* zusammen.[71] Sie weisen jedoch unmittelbar darauf hin, dass diese in der Praxis unerreichbar ist, da sich in realistischen Datensätzen gleiche Vorhersagegenauigkeit und gleiche Gruppengenauigkeit regelmäßig gegenseitig ausschließen. Je nach Sachlage und den gewünschten Ergebnissen können daher relevante Fairnesskonzepte angewendet und durch Algorithmen auf verschiedene Art und Weise sichergestellt werden. Es bedarf also einer Abwägung von Experten im Hinblick darauf, was in der jeweiligen Entscheidungssituation als Fairness bzw. Gleichbehandlung erstrebenswert ist, wie beispielsweise ob ein ADM-System kreditwürdige Männer und Frauen mit der gleichen Wahrscheinlichkeit als fälschlicherweise unwürdig einstufen sollte oder stattdessen die Wahrscheinlichkeit eines Kreditausfalls aufgrund einer falschen positiven Einstufung in beiden Gruppen gleich sein soll.

[71] Berk et al. 2017.

Beispiel Kreditscoring: Wir veranschaulichen die Nutzungsmöglichkeiten der erklärten Metriken anhand eines fiktiven Datensatzes und Klassifikationsmodells zur Kreditwürdigkeit von jeweils 200 Männern (M) und Frauen (F), wie sie beispielsweise anhand historischer Daten einer Bank erstellt werden könnten:

	Vorhersage Kreditwürdigkeit positiv	Vorhersage Kreditwürdigkeit negativ	Bedingte Gruppengenauigkeit
Tatsächlich kreditwürdig	*true positive* **M: 100** **F: 110**	*false negative* **M: 10** **F: 20**	*false negative rate* **M: 10 / 110 = 0,09** **F: 20 / 130 = 0,15**
Tatsächlich nicht kreditwürdig	*false positive* **M: 20** **F: 10**	*true negative* **M: 70** **F: 60**	*false positive rate* **M: 20 / 90 = 0,22** **F: 10 / 70 = 0,14**
Bedingte Vorhersage-genauigkeit	*positive prediction error* **M: 20 / 120 = 0,17** **F: 10 / 120 = 0,08**	*negative prediction error* **M: 10 / 80 = 0,125** **F: 20 / 80 = 0,25**	*accuracy (Genauigkeit)* **M: 170 / 200 = 0,85** **F: 170 / 200 = 0,85**

Zunächst zur zeilenweisen Betrachtung dieser Confusion Matrix, also der bedingten Gruppengenauigkeit: Von den 200 Männern im Datensatz sind tatsächlich 110 kreditwürdig und 90 nicht kreditwürdig. Von den 110 kreditwürdigen Männern hat das Modell 100 korrekt als solche identifiziert, jedoch 10 fälschlicherweise als nicht kreditwürdig eingestuft. Es hat somit eine *false negative rate*, also die Wahrscheinlichkeit einer inkorrekten Kreditverweigerung für kreditwürdige männliche Antragsteller, von 0,09. Umgekehrt klassifizierte das Modell 70 von 90 nicht kreditwürdigen Männern korrekt, würde aber entsprechend 20 Kreditausfälle verursachen. Dies ergibt eine *false positive rate*, also die Wahrscheinlichkeit einer inkorrekten Kreditvergabe an nicht kreditwürdige männliche Personen, von 0,22. Analog errechnen sich die entsprechenden Werte für weibliche Antragsteller. Es ist zu bemerken, dass der Anteil zu Unrecht abgelehnter kreditwürdiger Frauen im Vergleich fast das Doppelte beträgt (*false negative rate* von 0,15 gegenüber 0,09). Ferner ist die Wahrscheinlichkeit, trotz fehlender Kreditwürdigkeit ein Darlehen zu bekommen, für Männer ebenfalls deutlich höher als für Frauen (*false positive rate* von 0,22 gegenüber 0,14).

Die spaltenweise Betrachtung der bedingten Vorhersagegenauigkeit ergibt Folgendes: Von 120 Männern, denen das System einen Kredit gewähren würde, könnten 20 ihn nicht zurückzahlen. Der *bedingte Vorhersagefehler*, also die Wahrscheinlichkeit, dass ein an einen Mann gewährter Kredit ausfällt, liegt daher bei 0,17. Von den 80 abgelehnten männlichen Antragstellern hätten allerdings 10 das Darlehen zurückzahlen können. Somit liegt der *negative prediction error*, also die Wahrscheinlichkeit, dass ein Kredit zu Unrecht verweigert wurde, bei 0,125. Die Statistiken für Frauen ergeben sich analog. Es zeigt sich, dass der Anteil an Männer gewährten Krediten, die nicht zurückbezahlt werden, doppelt so hoch ist wie bei Frauen (*positive prediction error* von 0,17 gegenüber 0,08). Ferner ist der

Anteil der Frauen, denen zu Unrecht ein Kredit verweigert wurde, doppelt so hoch wie bei Männern (*negative prediction error* von 0,25 gegenüber 0,125).

Wenn nicht zwischen positiven und negativen Entscheidungen unterschieden wird, liegt die Gesamtgenauigkeit (*accuracy*, also der Anteil korrekter Vorhersagen an der Menge der Gesamtentscheidungen) des Modells bei 0,85 für Männer wie auch für Frauen. Wir erkennen, dass die Genauigkeit des Systems für beide Gruppen gleich ist, jedoch die Verteilung der Fehler des Systems zu Lasten der weiblichen Antragsteller geht.

Aus diesem Beispiel wird ersichtlich, dass das Vorhersageverhalten eines Klassifikationsmodells, insbesondere seiner Fehler, zwischen gleich zu behandelnden Gruppen mittels spezieller quantitativer Metriken sehr nuanciert untersucht werden kann. Ferner kann mittels dieser Metriken das Gebot der Gleichbehandlung präzisiert werden (siehe Fairnessbegriffe oben) und ADM-Verwender können ihr Modelltraining entsprechend auf die Einhaltung dieser Kriterien optimieren.

Wir präsentieren hier die Fairnesskonzeptionen von Berk et al. als eine mögliche Formalisierung von Fairnesskriterien in Klassifikationsproblemen. In der akademischen Forschung im Bereich *Fair Machine Learning* existieren zahlreiche andere Definitionen,[72] die teilweise äquivalent sind, sich überlappen, gegenseitig ergänzen oder ineinander überführbar sind. Ein weiterer, in der Literatur oft zitierter, quantitativer Ansatz für Unterschiedlichkeit und Gleichbehandlung nach Dwork et al. geht beispielsweise davon aus, dass es eine Funktion gibt, die die Ähnlichkeit zweier Datenpunkte (z.B. Individuen, die einen Kreditantrag stellen) ausschließlich anhand zulässiger Kriterien berechnet. Auf der Basis dieser Funktion lässt sich die Ungleichbehandlung von zwei Gruppen durch einen Klassifikator mit Hilfe der sogenannten *Lipschitz*-Bedingung begrenzen.[73] Nehmen wir eine solche Funktion an, die die „legitime“ Unterschiedlichkeit von beispielsweise zwei Kreditantragstellern in einer reellen Zahl quantifiziert. Aus Datenschutz- und Gleichbehandlungsimperativen folgt dann, dass bestimmte personenbezogene Attribute in diese Funktion entweder gar nicht oder nur unter bestimmten Bedingungen einfließen dürfen. Wenn bestimmte mathematische Kriterien erfüllt sind, lassen sich aus dieser Betrachtungsweise technische Konzepte zur Fairness eines Klassifikationsmodells ableiten, wobei der besagten Funktion zur Berechnung von Ähnlichkeit/Unterschiedlichkeit eine besondere Bedeutung zukommt und sie für jedes Sachproblem neu definiert werden muss.

Quantitative Fairnessbegriffe erlauben eine statistische Messung verschiedener Aspekte von Gleich- und Ungleichbehandlung von Gruppen durch ADM-Systeme, treffen jedoch typischerweise keine Aussage über die Struktur des verwendeten Modells oder die (nicht) verwendeten Variablen des Datensatzes. Sie sind somit komplementär zur Fairness-through-Blindness-Methode, nach der Ungleichbehandlungen dadurch vermieden werden sollen, dass die Gruppenzugehörigkeit vor dem lernenden Algorithmus möglichst effektiv maskiert wird, die jedoch das tatsächliche Verhalten (statistisches Verhältnis positiver/negativer Entscheidungen bzw. entsprechender Fehler in beide Richtungen) zwischen den Gruppen unter Umständen außer Acht lässt.

[72] Z.B. eine quantitative Definition von „Disparate Impact“ in Feldman et al. 2015; „Equalized Odds“ und „Equal Opportunity“ in Hardt et al. 2016; Corbett-Davies et al. 2017 mit weiteren Nachweisen zur Modellierung von Fairness, unter anderem bei *Risk Scores*.
[73] Dwork et al. 2012.

Das Problem der Verfügbarkeit von Testdaten: Wir führten ein Gespräch mit der Schufa Holding AG, welche Verbraucherdaten sammelt und durch ein Regressionsmodell erstellte Kreditwürdigkeitsbewertungen verkauft. Man erklärte uns, dass die entwickelten Regressionsmodelle geschützte Attribute nicht verwenden würden. Auf Nachfrage ergänzte man, dass eine statistische Überprüfung des trainierten Vorhersagemodells auf Ungleichbehandlung verschiedener Gruppen für das Geschlechtsattribut durchgeführt wurde. Hierbei wurde festgestellt, dass das Merkmal „Geschlecht" keine statistisch signifikante Korrelation zum finalen Score aufweise. Da Attribute wie „ethnische Herkunft" nicht erfasst werden (dürfen), könne auch keine diesbezügliche Untersuchung durchgeführt werden. Dies ist ein Beispiel für die praktische Relevanz von datensatzbasierten Gleichbehandlungsprüfungen. Zugleich weist es aber darauf hin, dass die Erstellung und Wartung von qualifizierten und die Realität repräsentativ abbildenden Testdatensätzen sehr schwierig sein kann.

Vor diesem Hintergrund sei auf ein Experiment in Berk et al. (2017) hingewiesen[74], in dem ein Klassifikationsmodell für Bewährungsentscheidungen entgegen dem Datenschutzgebot explizit durch Training zweier separate *Random Forest*-Modelle für weiße und afroamerikanische Individuen erstellt und auf Fairnessverhalten überprüft wurde. Ein solches Vorgehen dient zwar dazu, eine Diskriminierung zu verhindern und quantitative Gleichbehandlungsgebote zu erfüllen; gleichzeitig erfordert es aber die Verarbeitung besonderer Kategorien personenbezogener Daten und steht damit gegebenenfalls im Widerspruch zu europäischem Datenschutzrecht. Trotzdem bildet dieses Experiment vor dem Hintergrund quantitativer Fairnessbegriffe einen nützlichen Referenzpunkt in der Diskussion um die Geeignetheit und Effektivität des Datenschutzrechts zur Sicherstellung von Gleichbehandlung durch ADM-Systeme.

Die entscheidende Grundsatzfrage ist hier, wo Gleichbehandlung bzw. Fairness funktional ansetzen soll, also entweder bei der Nichtberücksichtigung der Gruppenzugehörigkeitsattribute (ex ante / prozessorientiert) und/oder durch empirische Kriterien (ex post / ergebnisorientiert).

4.3.3 Zwischenfazit

Es existieren quantitative Fairnessbegriffe, die nicht an der unzulässigen Verwendung der Attribute durch ein ADM-Modell ansetzen, sondern die empirisch-statistische Verteilung des Vorhersageverhaltens auf einem gesamten Datensatz prüfen. Hierbei lässt sich das Konzept der (Un-)Gleichbehandlung in verschiedene Aspekte aufgliedern, wie z.B. gleiches Verhältnis von positiven und negativen Entscheidungen in jeder Gruppe oder gleiche Irrtumswahrscheinlichkeit in allen Gruppen. Manche dieser Gleichheitsformen schließen sich in der Praxis gegenseitig aus, weshalb im jeweiligen Einzelfall von Experten bestimmt werden muss, welche Formen der Gleichheit zu priorisieren sind.

Wenn quantitative Fairnesskriterien als Maßstab für ADM-Systeme angenommen werden und entsprechende Testdaten verfügbar sind, lassen sich Prüfungen sehr transparent und effektiv gestalten. Quantitative Fairnessbegriffe sind ferner praktisch dadurch relevant, dass sie als Beschränkungen in Lernalgorithmen eingesetzt werden können, d.h., Algorithmen zum Maschinellen Lernen können quantitative Fairnesskriterien in ihren Optimierungsprozess (also den Lernprozess selbst) mit einbeziehen.

[74] Berk et al. 2017 wiederum verweisen auf vorangegangene Experimente in Žliobaitė 2016.

Eine empirische Untersuchung dieser theoretisch-prinzipiellen Geeignetheit in verschiedenen und teils sehr komplexen praktischen Anwendungsfeldern von ADM steht noch aus. Quantitative Fairness und ihre Berücksichtigung in ML-Modellen sind Gegenstand aktiver Forschung und unserer Ansicht nach daher gegenwärtig noch nicht reif zur Nutzung als prinzipielle Anforderung oder regulatorisches Instrument für ADM-Systeme. Sie bergen jedoch Potenzial für künftige Machbarkeitsstudien oder Forschungsförderung im Hinblick auf die Nutzung dieser Methoden in begrenzten Anwendungsgebieten, wie beispielsweise einer standardisierten Zertifizierung oder einem Auditing von Kreditscoringsystemen durch die Bankenaufsicht.

4.4 Kontrolle von ADM-Systemen

Algorithmische Entscheidungsfindung und die Erstellung von Systemen, die auf Basis von trainierten Algorithmen Entscheidungen treffen können, sind ein komplexer Prozess, der aus mehreren Einzelschritten besteht. Die Planung eines solchen Systems sowie dessen Umsetzung sind nicht trivial und es existieren zahlreiche Designentscheidungen, die die Struktur und auch das Verhalten eines solchen Systems beeinflussen. In aktuellen populärwissenschaftlichen Auseinandersetzungen mit dem Thema „Künstliche Intelligenz“ wird sehr häufig nur auf einen Teilbereich des Systems geblickt. Dies ist jedoch für den Entwurf und die Entscheidungen eines implementierten Systems zur algorithmischen Entscheidungsfindung nicht ausreichend.

Wie bereits erläutert wurde, existieren verschiedene Gruppen trainierbarer Modelle, die in Datensätzen vorhandene Muster, Regelmäßigkeiten und Anomalien erfassen und in einem mathematisch berechenbaren Modell formalisieren. Diese Lernprozesse unterscheiden sich zum Teil erheblich voneinander. Allerdings ist diese Formalisierung auch nur ein Teilschritt, der in den Gesamtprozess, in dem die algorithmische Entscheidungsfindung stattfindet, eingebettet ist. Deswegen ist es auch für eine wertende Betrachtung nicht ausreichend ausschließlich diese Erstellung eines Vorhersagemodells zu berücksichtigen.

Die Differenzierung und Analyse der Einbettung in den Gesamtprozess wurde in einem ähnlichen Kontext von anderen Autoren bereits vorgeschlagen und scheint zielführend für die Analyse von ADM-Systemen[75]. Krüger und Lischka weisen zu Recht auf die Komplexität des Feldes hin und erarbeiten in einer sorgfältigen Begriffsdifferenzierung des Feldes ADM einen Beitrag zur Bewältigung der Herausforderungen.

Es können drei unterschiedliche Ebenen zur Überprüfung und zur Analyse der Transparenz unterschieden werden[76]:

1. Analyse des Gesamtprozesses zur (Weiter-)Entwicklung eines ADM-Systems
2. Analyse des trainierten Modells (als *Whitebox* oder *Blackbox*)
3. Analyse der Entscheidung auf Instanzebene

Unter Berücksichtigung dieser drei Ebenen kann es gelingen, ein ADM-System zu analysieren und einen holistischen Überblick über seine Eigenschaften zu bekommen. Diese sind nicht notwendigerweise statisch und können sich über den Lebenszyklus des Systems

[75] Krüger/Lischka 2018.
[76] Waltl/Vogl 2018.

verändern. Das ist beispielsweise der Fall, wenn ein bestehendes Modell unter Berücksichtigung neuer Daten weiter trainiert wird. Die Erfassung dieser Umstände ist unter Verwendung des hier vorgestellten Modells vorgesehen.

4.4.1 Analyse des Gesamtprozesses

Der Entwurf und die Implementierung eines ADM-Systems sind in einen Prozess eingebettet, der das Training möglicher ML-Modelle als einen Teilschritt umfasst. Die verbleibenden Schritte können vereinfacht wie in Abbildung 2 gezeigt zusammengefasst werden. Die Abbildung wurde von Waltl und Vogl (2018) umfassend diskutiert und orientiert sich an dem Standardprozess für KDD (Knowledge Discovery in Databases) von Fayyad (1996).

Abbildung 2: Fünf Schritte des Standardprozesses für Knowledge Discovery in Databases

Der Prozess umfasst fünf Schritte, die unmittelbar aufeinander folgen:

1. *Data acquisition*: Zunächst müssen Daten erhoben und gespeichert werden, auf deren Basis die Entscheidungsstrukturen trainiert und optimiert werden. Im weitesten Sinne zählt hierzu jede Technologie, die angewandt wird, um (verbraucherrelevante) Daten zu erheben. Auch das Konsolidieren von Daten aus unterschiedlichen Quellen wird innerhalb des *Data-acquisition*-Schritts durchgeführt. Das Konsolidieren ist deshalb notwendig, weil Daten über Personen von unterschiedlichen Quellen, z.B. Familiensituation, Einkaufsverhalten, Zahlungsbereitschaft und Zahlungsausfälle sowie Freizeitinteressen etc., zusammengeführt werden können. Diese müssen nicht notwendigerweise alle über das Internet akquiriert werden, sondern können auch von analogen Quellen gespeist werden. In Fachkreisen zählt man hierzu auch das komplexe Forschungsgebiet des sogenannten *User Modelings*[77]. Hierbei geht es darum, eine digitale Repräsentation von Benutzern zu erhalten und möglichst zielgerichtet Information über diese zu sammeln. Dies ist die Grundlage für nachfolgende Analyseverfahren. Es zeigt sich, dass diese bereits wegweisend für die Entscheidungen sein können, die nachfolgend von Algorithmen getroffen werden. Im wissenschaftlichen Feld des „*discrimination-aware data mining*“ werden aktiv Probleme bearbeitet, die sich ergeben, wenn bereits bei der Auswahl und Erhebung der Daten Biases (und Vorurteile) miterhoben werden[78]. Die nachgelagerte mathematische Analyse könnte genau diese Unausgewogenheiten in den Daten erkennen und in den algorithmisch getroffenen Entscheidungen berücksichtigen.

2. *Pre-processing*: Die erhobenen Daten werden im nächsten Schritt verarbeitet und derart aufbereitet, dass fehlende oder nicht verifizierbare Informationselemente über eine einzelne Person zu handhaben sind (z.B. fehlende oder potenziell unrichtige

[77] Kobsa, Alfred. "Generic user modeling systems." *User modeling and user-adapted interaction* 11.1-2 (2001): 49-63.
[78] Pedreshi et al. 2008.

Angaben zur Anzahl von Kindern). Der Datenpunkt kann dann beispielsweise ignoriert, durch einem Standardwert ersetzt oder mit einer anderen Methode durch einen künstlichen Wert substituiert werden. Das *Pre-processing* bestimmt entscheidend über die zur Verfügung stehenden Daten mit und kann erheblichen Einfluss auf das Modelltraining und folglich auf die Qualität der algorithmischen Entscheidung haben. Daten werden daher typischerweise so gespeichert, dass zum Zeitpunkt des Modelltrainings verschiedene Formen von *Pre-processing* möglich sind.

3. Transformation: Hierin werden Informationselemente in bestimmte Datentypen vereinheitlicht. Zahlenwerte werden in vergleichbare Skalen gebracht (z.B. Saldierung verschiedener Einkommensangaben), Textinformationen werden in eine abschließende Menge von Kategorien eingeordnet (z.B. Normalisierung von Freitextangaben zum Beruf) und kategorische sowie binäre Informationen entsprechend erfasst (z.B. Familienstand, Ausbildung). Falls erforderlich und sinnvoll werden metrische Datumswerte in eine ordinalskalierte Ordnung übertragen. Dies könnte z.B. durch die Überführung von Eigenschaften einer Person wie Größe oder Alter in Klassen stattfinden, wie beispielsweise sehr groß, groß, mittelgroß, klein, sehr klein usw. Im Rahmen dieser Kategorisierung und Klassifizierung modifiziert man den Datenbestand derart, dass man einige bestimmte Phänomene verstärken bzw. abschwächen kann. Zahlenwerte können beispielsweise entweder direkt in die Repräsentation fließen oder vorher in eine Zahl zwischen 0 und 1 mittels eines Minimal- und Maximalwerts projiziert werden, was potenziell Konsequenzen für das spätere Modelltraining haben kann.[79]

4. *Training & application of AI model*: Auf Basis der erhobenen, bereinigten und transformierten Daten werden anschließend die Modelle trainiert. Das Training verläuft dabei ganz unterschiedlich und variiert stark nach dem eingesetzten Modelltyp (siehe Kapitel 4.1 „Grundlagen Maschinellen Lernens“). Grundlegende Funktionsweise ist das Optimieren von mathematischen Funktionen und Operationen zur Approximation der Trainingsdaten. Das bedeutet, dass bestehende Methoden aus Bereichen der Wahrscheinlichkeitstheorie oder Statistik, aber auch Informationstheorie verwendet werden, um die Zusammenhänge zwischen den Attributen in den aufbereiteten Trainingsdaten bestmöglich beschreiben zu können. Bestmöglich in diesem Zusammenhang bedeutet in der Regel fast immer, dass man die Parameter zu einer bestehenden Funktion derart wählt, dass eine festgelegte Fehlerrate (*Loss Function*) möglichst klein ist. Wie in Kapitel 4.1 erklärt ist das Ergebnis des Trainingsprozesses ein Vorhersagemodell bestehend aus einem Algorithmus zur Vorhersage (z.B. einer mathematischen Gleichung) sowie einer Menge „gelernter“ Parameter.

Außerhalb der Trainingsphase werden die Parameter normalerweise nicht mehr verändert, so dass das Entscheidungsverhalten festgelegt und nachvollziehbar ist. Da die zugrundeliegenden mathematischen Operationen in der Regel auch nicht auf Zufallsvariablen basieren, sind die Modelle deterministisch und das Entscheidungsverhalten entspricht exakt reproduzierbar der trainierten und parametrisierten Funktion.[80] Eine Entscheidung bedeutet in diesem technischen Sinne,

[79] Diese Projektion ist eine Form des sogenannten „Feature Scaling“, welches auch andere Formen annehmen kann.

[80] Im Gegensatz zur Vorhersage durch das trainierte Modell können im Training selbst durchaus Zufallselemente eine Rolle spielen. Auf den gleichen Trainingsdaten können mit dem gleichen Trainingsalgorithmus verschiedene Modelle produziert werden. Typische Zufallselemente im Trainingsprozess sind beispielsweise die anfängliche Initialisierung von Modellparametern vor der Optimierung oder stochastisch beeinflusste Optimierungsvorgänge wie etwa Simulated Annealing. In

dass bestimmte Eingabewerte an diese Funktion weitergereicht werden, die dann eine Vorhersage berechnet und dabei Parameter verwendet, welche auf der Basis des Trainingsdatensatzes optimiert wurden. An dieser Stelle wird nochmals offensichtlich, dass die ausgewählten Trainingsdaten (*Data acquisition* und *Pre-processing*) maßgeblich darüber entscheiden, wie Entscheidungen über noch unbekannte (neue) Daten getroffen werden.

5. *Post-processing:* Im Anschluss an das Trainieren und Auswerten der Funktion in Bezug auf einen neuen Datensatz kann der Ausgabewert der Funktion unter Umständen noch nachbearbeitet werden. Dies ist insbesondere dann der Fall, wenn eine Funktion einen metrischen Wert zurückliefert (z.B. bei Regressionen), die finale Entscheidung aber eine binäre Entscheidung (Ja/Nein) sein soll. Hier können anstatt einer logistischen Funktion (siehe 4.1.1) auch Schwellenwerte angesetzt werden, die entsprechend den Ausgang einer Entscheidung bestimmen oder empfehlen. Viele Implementierungen liefern einen Zahlenwert zu ihrem Ergebnis, der darüber Auskunft gibt, wie „sicher" (Konfidenz) sich das Modell bei dieser Entscheidung war. Veranschaulichen kann man es sich an einem vereinfachten Beispiel: Wenn eine Entscheidung zu einem Datensatz getroffen werden soll, der so oder ganz ähnlich sehr häufig in den Trainingsdaten auftaucht und dort zu immer der gleichen Entscheidung geführt hat, so kann über diesen neuen Datensatz normalerweise auch *confident* entschieden werden. Sind die Attribute aber sehr verschieden von jenen innerhalb der Trainingsdaten, oder sind über diese Ausprägung der Attribute nur sehr wenige Datenpunkte verfügbar, so sinkt die Konfidenz und auch die Zuverlässigkeit des Ergebnisses kann abnehmen. Verschiedene ML-Modelle haben unterschiedliche Methoden und Fähigkeiten, um mit in den Trainingsdaten mehr oder weniger dicht abgedeckten Bereichen des Attributraums umzugehen.

Die oben genannten Überlegungen betten das ADM in einen sehr komplexen Prozess ein, in dem jeder Schritt mitentscheidend für das schlussendliche Entscheidungsverhalten ist. Der Prozess entspricht einem Standard von 1996 und ist wasserfallartig angelegt. Das bedeutet, dass er Schritt für Schritt durchlaufen wird. In der Praxis laufen diese Prozesse komplexer ab, da noch zusätzliche Iterationen verwendet werden. Als ein etablierter Industriestandard hat sich der *Cross-industry standard process for data mining* (CRISP-DM) durchgesetzt[81].

Praxis: Wir führten ein Gespräch mit der Schufa Holding AG, welche Verbraucherdaten sammelt und Bonitätsprüfungen (*Kreditscoring*) mit ML-Verfahren durchführt. Der dort stattfindende Prozess lässt sich mit der hier vorgestellten Methode abbilden.

1. *Data acquisition:* Daten über Kreditnehmer werden erhoben, von Geschäftspartnern geliefert oder anderweitig besorgt. Umfangreiche Datenbanken über Personen werden gepflegt. Sie stellen die Grundlage für die Analysen und Vorhersagemodelle (logistische Regression) dar.

der Praxis wird dies mit Hilfe von Programmen zur zufälligen Generierung von Zahlen nach bestimmten mathematischen Verfahren implementiert. Diese wiederum lassen sich optional durch einen Startwert so parametrisieren, dass sie reproduzierbar die gleiche Sequenz an Zufallszahlen generieren. Zufall kann im Modelltraining also eine Rolle spielen, jedoch existieren Methoden, um dennoch, falls gewünscht, Reproduzierbarkeit im Trainingsprozess herzustellen.

[81] Kurgan/Musilek 2006; Mariscal et al. 2018.

2. *Pre-processing:* Die gesammelten Daten über Personen werden aufbereitet. Fehlende Datumswerte werden ergänzt und Ausreißer bzw. fehlerhafte Daten müssen erkannt und entfernt werden, um das Ergebnis nicht zu verfälschen.

3. *Transformation:* Daten aus unterschiedlichen Quellen müssen unter Umständen auch harmonisiert und vereinheitlicht werden. Hierzu zählt auch das Aggregieren von Attributen, z.B. um die durchschnittliche Zahlungsdauer auf Basis von vielen Zahlungen zu berechnen.

4. *Training & application of AI model:* Die Attribute einer Person werden in einem ML-Verfahren, in diesem Fall einer logistischen Regression, verarbeitet. Das Verfahren besteht aus einem Algorithmus, der im Kern eine mathematische Operation ausführt. Die Gewichte der jeweiligen Attribute werden in der Trainingsphase bereits ermittelt. Diese werden in sogenannten Score Cards repräsentiert. Auf Basis des Inputs, der gesammelten Daten einer Person, wird ein Output, die Ausfallwahrscheinlichkeit, berechnet.

5. *Post-processing:* Die Ausfallwahrscheinlichkeit wird einem Geschäftspartner zur Verfügung gestellt, der darauf aufbauend eine Entscheidung über die Kreditwürdigkeit einer Person trifft. Der Score wird damit zu einer wichtigen Variablen innerhalb des Entscheidungsprozesses.

Dieser Prozess ist in der untenstehenden Abbildung dargestellt. Noch stärker als der soeben erläuterte KDD-Prozess bezieht der CRISP-DM Anforderungen aus dem Geschäftsfeld mit ein, die ganz zu Beginn des Prozesses erhoben werden („*Business Understanding*“). Anschließend werden die verfügbaren Daten analysiert und gegebenenfalls noch erhoben („*Data Understanding*“). Hierin zeigt sich auch eine Schwäche des CRISP-DM gegenüber dem oben diskutierten Vorgehensmodell: Die Erhebung der Daten („*Acquisition*“) wird nicht mehr explizit berücksichtigt, sondern implizit angenommen oder als eine vorgelagerte Operation verstanden, was der Komplexität des Gesamtproblems nicht gerecht wird. Hier können sich zwei Probleme ergeben. Zum einen können die erhobenen Daten nicht repräsentativ für die praktische Anwendung sein, wodurch die Wahrscheinlichkeit von inhaltlich unrichtigen Entscheidungen möglicherweise erhöht wird. Zum anderen kann die Erhebung unausgewogene Daten produzieren, welche zu einem unausgewogenen Entscheidungsverhalten führen können. Repräsentativität und Unausgewogenheit (*Biases*) hängen zwar oft zusammen, sind aber prinzipiell getrennte Probleme. Repräsentative Daten können durchaus unausgewogen sein, was in der Gestaltung des Lernverfahrens zu berücksichtigen ist. Es ergibt sich daher, dass Datenerhebung und Modelltraining stark zusammenhängen und ein ADM-Prozessmodell beide Aspekte und ihre Interaktion berücksichtigen sollte (vgl. Kapitel 4.2).

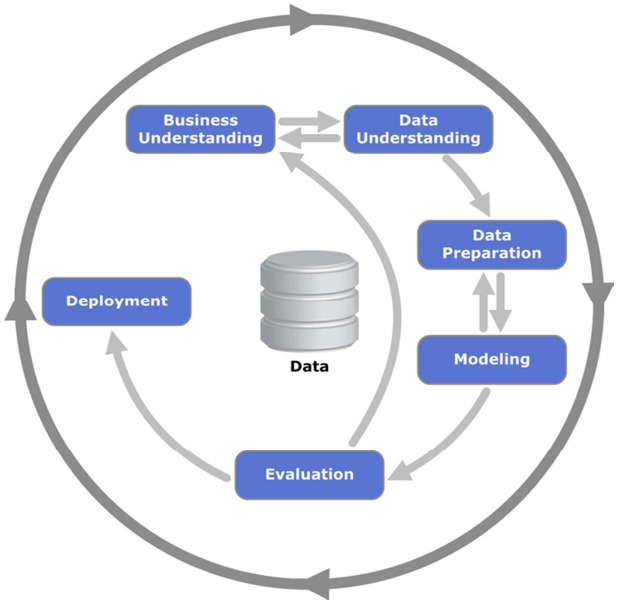


Abbildung 3: Cross-industry standard process for data mining[82]

Der CRISP-DM stellt den industriellen Einsatz von Datenanalyse und *Machine Learning* stärker in den Vordergrund. Dies zeigt sich insbesondere durch die Abbildung der Aktivitäten „*Business Understanding*“, „*Data Understanding*“ und „*Deployment*“. Für die Praxis stellen die Aktivitäten wichtige Schritte bei der Entwicklung von ADM-Systemen dar. Bei der Detektion und zur Vermeidung von nicht ausgewogenem Vorhersageverhalten von ADM-Systemen spielen diese indes keine bzw. nur eine untergeordnete Rolle. Die Vermeidung von unzulässiger Ungleichbehandlung kann jedoch schon sehr früh in dem Prozess als eine funktionale Anforderung erklärt und bei der Aktivität „Data Understanding“ berücksichtigt werden. Sinnvolle Fragen könnten darauf abzielen, ob die vorliegenden Daten repräsentativ sind und ob alle Klassen und Attribute ausreichend abgebildet sind, um einen zu starken Bias bei der nachfolgenden Modellierung zu vermeiden (vgl. Kapitel 4.2.1). Die beiden Phasen „*Data Preparation*“ und „*Modeling*“ beinhalten den DM-Prozess von Fayyad (siehe oben). Diese stehen auch im CRISP-DM in einem wechselseitigen Abhängigkeitsverhältnis und bedingen sich gegenseitig.

Die Erläuterungen zeigen, dass es möglich ist, Fragen nach Ungleichbehandlung und Fairness im Prozess der Erstellung von ADM-Systemen zu berücksichtigen. Insbesondere in den frühen Phasen der Systemerstellung, Datenerhebung und Datenaufbereitung ist darauf zu achten, eine systematische Ungleichbehandlung zu vermeiden. Grundvoraussetzung dafür ist eine Sensibilisierung für eine Ungleichbehandlung durch Algorithmen. Hinzu kommen Audit- und Testverfahren, die verwendet werden können, um das Verhalten eines ADM-Systems exakt zu beschreiben (siehe Kapitel 4.4). Eine Konkretisierung in Hinblick auf Fragestellungen des Kreditscorings bzw. verbraucherschutzrechtlicher Systeme bleibt hier noch offen, da es kaum möglich ist pauschale Antworten ohne genaue Kenntnis der Systeme und der zugrundeliegenden Anforderungen („Business Understanding“) zu geben. Auf Basis der Überlegungen zur Analyse und Differenzierung des ADM-Prozesses werden im nächsten Kapitel die Möglichkeiten zur Analyse des trainierten Modells betrachtet. Dies betrifft vor allem den vierten Schritt *(„Training & application of AI Model“)* in dem oben vorgestellten Prozess von Fayyad.

[82] Illustration von Kenneth Jensen [CC BY-SA 3.0], via Wikimedia Commons.

4.4.2 Technische Analyse von Machine-Learning-Modellen

Im folgenden Kapitel wird auf die Möglichkeiten, aber auch die Herausforderungen der Analyse von Verfahren zur algorithmischen Entscheidungsfindung (ADM) auf Basis von Maschinellem Lernen eingegangen. Hierbei wird insbesondere zwischen Whitebox- und Blackbox-Sichten unterschieden. Außerdem wird anhand anschaulicher Beispiele illustriert, welche Methoden zur Sichtbarmachung von Entscheidungsstrukturen verwendet werden können und welche Herausforderungen sich ergeben.

4.4.2.1 Analyse eines Whitebox-Modells ohne Daten

4.4.2.1.1 Erklärung von Modellen in ADM-Systemen

Dass ein System, das automatisch und auf Basis von trainierten mathematischen Modellen Entscheidungen trifft diese anschließend erklären soll, ist keinesfalls selbstverständlich. In anderen Disziplinen, insbesondere in der früheren Forschung im Bereich Künstlicher Intelligenz, wurde dieser Anspruch schon früher gestellt. Bei sogenannten Expertensystemen, also in Systemen, in denen Fachwissen explizit modelliert und formalisiert wird, wird die Erklärungskomponente als ein zentraler Bestandteil des Gesamtsystems betrachtet und trägt wesentlich zu seinem Nutzungswert bei. Für ADM unter Benutzung von statistischen bzw. ML-Modellen muss der Brückenschlag zwischen Erklärung und Technologie aber erst noch grundlegender verstanden und beschrieben werden. Hilfreich scheinen hier Leitfragen zu sein, anhand derer die notwendigen Richtungen vorgegeben werden, hinsichtlich derer die Rolle von Erklärungen differenziert betrachtet und der Einfluss auf Technologie konkreter beschrieben werden kann.

Ein konstruktiver Vorschlag für Leitfragen wurde beispielsweise von David Gunning (2017) im Rahmen einer Präsentation von *Explainable Artificial Intelligence (XAI)* durchgeführt. Im Folgenden benutzen wir diese als Grundlage für die Formulierung von konkreten Fragen, die es erlauben die technische Machbarkeit eines Algorithmengesetzes und die Regulierung von ADM zu prüfen:

1. Warum hat das ADM eine bestimmte und keine andere Entscheidung getroffen?

Frage 1 zielt auf das konkrete Entscheidungsverhalten eines ADM ab und legt das Hauptaugenmerk auf die getroffene Entscheidung im Verhältnis zur Menge der potenziell möglichen Entscheidungen. Diese Menge aller möglichen Entscheidungen lässt sich in der Regel nur vor dem Hintergrund des Sachproblems ermitteln, dem das ADM-System dienen soll. Für Fragestellungen der Bonitätsprüfung in dem Bereich Kreditscoring wird üblicherweise ein Zahlenwert berechnet, der die Kreditwürdigkeit einer Person repräsentiert. An einer Stelle des Entscheidungsprozesses wird dieser Zahlenwert (Score) auf eine Ja-oder-Nein-Entscheidung reduziert. Die Frage, ob jemand einen Kredit bekommt oder nicht, hängt damit oftmals unmittelbar mit dem Credit-Score zusammen, auch wenn für die endgültige Entscheidung noch zusätzliche Faktoren berücksichtigt werden. Um zu erläutern, warum eine Entscheidung getroffen wurde, ist es deshalb notwendig zu klären, welche Entscheidungsmöglichkeiten ein ADM-System hat und welche Faktoren während des Entscheidungsprozesses berücksichtigt werden.

2. Für welche Menge von Entscheidungen funktioniert das ADM (nicht)?

Um die Nutzungspotenziale und Grenzen algorithmischer Entscheidungsfindung in einem Sachproblem einzuschätzen und konstruktiv zu beschreiben, gilt es die Menge von Fällen

einzugrenzen, für die ein entwickeltes ADM-System richtige Entscheidungen treffen kann. Diese Grenzen ergeben sich aus vielerlei Faktoren. Eine Grundannahme im Gebiet des Maschinellen Lernens ist etwa, dass die Verteilung der Testdaten (d.h. die in der Benutzung eingegebenen Daten zur Entscheidung) weitestgehend der Verteilung der Trainingsdaten folgt. Ein Kreditscoringsystem, das anhand von Daten einer bestimmten Region trainiert wurde, kann in einer anderen Region möglicherweise nicht ohne Anpassung angewendet werden. Auch sollten die Daten strukturell mit dem Modell kompatibel sein und keine unerwarteten Lücken, Falschinformationen oder Messfehler enthalten.

Die schwierigste Abschätzung in diesem Zusammenhang ist die Identifikation von Bereichen des Attributraumes, in dem das Modell ungenau ist, entweder weil es zu wenig Trainingsbeispiele in diesem Bereich zur Verfügung hatte, das Modell technisch zu limitiert ist, die Phänomene in den Daten zu erfassen, oder weil die Trainingsdaten selbst unsauber oder inkohärent sind. Idealerweise sollte ein ADM-System derart konstruiert sein, dass es zusätzlich zur getroffenen Entscheidung auch noch Auskunft über die Konfidenz geben kann, also über die eigene „Zuversicht“ in die Richtigkeit der Vorhersage.[83] Dies erlaubt die Unterscheidung in eine Menge von Eingaben, für die das System eine richtige Entscheidung trifft und sich dabei auch sehr sicher ist (hohe Konfidenz), und solche Eingaben, für die das System erkennt, dass es auf Basis der trainierten Entscheidungsstrukturen nicht mit hoher Konfidenz auf richtige Ergebnisse schließen kann. Letzteres erlaubt beispielsweise manuelle Kontrollmechanismen.

Wichtig ist hier die Unterscheidung zwischen richtigen Ergebnissen, Ergebnissen mit hoher Konfidenz und nichtdiskriminierenden Ergebnissen. Diese drei Charakteristika richtig/falsch, hohe/niedrige Konfidenz und diskriminierend/nichtdiskriminierend hängen nicht notwendigerweise voneinander ab und beschreiben unterschiedliche Aspekte einer Entscheidung. So kann eine „richtige“ Entscheidung aus Sicht des maschinell trainierten Systems getroffen werden, die dennoch diskriminierend ist. Ohne ein geeignetes Maß zur Bewertung von Diskriminierung spiegelt das System nur den Trainingszustand und die mathematischen Funktionen (inklusive Parameter) wider, auf die es trainiert wurde. Analog kann es auch passieren, dass eine falsche Entscheidung mit einer hohen Konfidenz getroffen wurde. Das System ist sich also sehr sicher bei seiner Entscheidung, trifft aber nach menschlichen Maßstäben dennoch die falsche Entscheidung. Zusätzliche Mechanismen und Methoden sind somit erforderlich, um das Entscheidungsverhalten von ADM bewerten und Fragestellungen der Diskriminierung beantworten zu können.

3. Auf Basis welcher Daten, Features (inklusive Gewichtung) und mathematischer Operationen wurde die Entscheidung getroffen?

Frage 3 konkretisiert die Fragen 1 und 2 weiter. Üblicherweise benötigen ADM-Systeme zur Entscheidungsfindung Eingabedaten, die von außen an sie herangetragen werden. Diese werden dann vorverarbeitet und durchlaufen diverse Verarbeitungs- und Transformationsschritte, bis sie der Komponente zum Maschinellen Lernen übergeben

[83] Eine einfache technische Implementierung dieses Prinzips im Bereich Kreditscoring kann sein, dass das System nicht nur ausgibt, ob die Bewertung des Antragstellers den Schwellenwert zur positiven Entscheidung überschreitet, sondern auch den relativen Abstand zur Schwelle anzeigt. Effektiv läuft dies auf eine prozentuale Charakterisierung des Ausfallrisikos hinaus, welche wiederum als Angabe in übergeordnete quantitative Entscheidungsprozesse einfließen kann (z.B. unter Hinzunahme des potenziellen Gewinns). Aus unseren Gesprächen mit Nutzern der Technologie haben wir Grund zur Annahme, dass diese Verwendung von ML-Modellen in der Praxis zumindest im Bereich Kreditscoring üblich ist.

werden, die auf Basis der trainierten internen Entscheidungsstruktur eine Entscheidung herbeiführt. Diese interne Entscheidungsstruktur berücksichtigt nicht notwendigerweise jedes zur Verfügung gestellte Attribut. Zusätzlich können unterschiedliche Attribute mit gänzlich unterschiedlichen Gewichten zu einer Entscheidung beitragen. Deshalb trägt Frage 3 dazu bei, dass dieser Umstand bei der Analyse von ADM-Systemen berücksichtigt wird. Vor diesem Hintergrund tauchte eine zusätzliche Facette auf, die bei der Frage nach Diskriminierung berücksichtigt werden muss. Die bloße Formulierung der Frage bedeutete noch nicht, dass diese auch einfach zu beantworten ist. Auf die Herausforderung, die Gewichte und die interne Entscheidungsstruktur offenzulegen, wird im Kapitel „Analyse von ADM“ noch explizit hingewiesen.

Als Ergänzung zu Frage 2 adressiert Frage 3 die Verwendung von mathematischen Operationen und Methoden innerhalb eines ADM-Systems. Während in der vorherigen Frage insbesondere die Verwendung von Daten, Attributen und deren Gewicht im Vordergrund steht, zielt diese Leitfrage darauf ab die Zusammenhänge zwischen diesen Attributen zu analysieren. Wie bereits gezeigt wurde, existieren zahlreiche unterschiedliche Ansätze, die für maschinelles Lernen verwendet werden können.

4.4.2.1.2 Beispiele von Techniken erklärbarer Modelle

Logistische Regression

Ein weit verbreitetes Vorhersagemodell zur Klassifikation ist die in Kapitel 4.1.1 bereits erklärte logistische Regression. Hier werden die Eingabevariablen mit statistisch gelernten Koeffizienten-Parametern multipliziert, aufsummiert und in den Bereich zwischen 0 und 1 projiziert.

Zur Erinnerung: Wenn $\hat{\mathbf{y}}$ die Zielvariable ist (z.B. die Kreditwürdigkeit einer Person) und $\mathbf{x_1}$, …, $\mathbf{x_n}$ die Eingabevariablen sind, dann ergibt sich das Modell aus folgender Gleichung:

$$\hat{y} = \sigma(w_1 x_1 + w_2 x_2 + ... + w_n x_n)$$

Die Koeffizienten w_1, …, w_n sind Modellparameter, deren genauer Wert aus den Daten gelernt wird. Das griechische Symbol **σ** ist hierbei eine sogenannte „logistische Funktion“, die eine beliebig große oder kleine Zahl in das Intervall zwischen 0 und 1 projiziert, um die Kreditwürdigkeit als binäre Entscheidung abzubilden.[84] Beispielsweise wird das Alter des Antragstellers als natürliche Zahl, das Jahreseinkommen in tausend Euro als natürliche Zahl, die Anzahl der minderjährigen Nachkommen als natürliche Zahl sowie der Ehestatus als 0 (ledig) oder 1 (verheiratet) jeweils mit einem solchen Parameter multipliziert, das Ergebnis aufsummiert und schließlich auf einen Wert zwischen 0 und 1 reduziert.

Das Problem ist entsprechend, die richtigen Parameter zu finden, so dass die Regressionsgleichung für kreditunwürdige Antragsteller einen Vorhersagewert möglichst nah an 1 und für kreditunwürdige möglichst nah an 0 ergibt. Trainingsdaten vorausgesetzt, lassen sich diese Parameter (also ein *Regressionsmodell*) von Daten mit Hilfe eines Optimierungsvorgangs lernen. Im Anschluss an den Lernvorgang können nun die Parameter jeder einzelnen Variablen untersucht werden. Ist ihr absoluter Wert sehr klein, so hat das

[84] In diesem Beispiel handelt es sich um eine „Sigmoid-Funktion“, die beliebige Zahlen in das (0, 1)-Intervall abbildet. Sie sind monoton steigend und nähern sich für sehr kleine Werte an die Untergrenze und für sehr große Werte an die Obergrenze des Intervalls an. Die Zahl 0 wird dementsprechend auf den Mittelpunkt des Intervalls (z.B. 0,5) projiziert.

entsprechende Attribut nur wenig bis keinen Einfluss auf die Voraussage. Ist jedoch der Parameter weit von 0 entfernt (positiv wie negativ), so hat das Attribut entsprechend Einfluss.

Als illustratives Beispiel für eine solche Untersuchung von gelernten Koeffizienten und ihrer Konsequenzen im Bereich der Fair-Lending-Aufsicht in den USA verweisen wir an dieser Stelle auf einen Artikel von Calem und Longhofer[85], dessen detaillierte Tabellen von Attributen und Koeffizienten wir hier aus Platzgründen nicht reproduzieren können.

Wir halten fest dass die logistische Regression kein Blackbox-Modell ist, da ihre Parameter nach dem Lernprozess einfach einseh- und interpretierbar sind. Unter anderem deswegen ist sie nach wie vor ein sehr verbreitetes Modell in allen Bereichen, in denen Interpretierbarkeit und Transparenz wichtig sind und die Einbuße an Genauigkeit gegenüber komplexeren Modellen verschmerzbar ist.

Entscheidungsbäume

Ein weiteres gut interpretierbares Modell sind sogenannte Entscheidungsbäume (*Decision Trees*). Hierbei wird die Voraussage durch die Verzweigung von Attributprüfungen vorgenommen. Ein solcher Entscheidungsbaum kann ebenfalls durch Daten trainiert/gelernt werden[86] und bildet im einfachen Fall das gesamte Modell.[87] Entscheidungsbäume sind der logistischen Regression in vielerlei Hinsicht überlegen. Wenn beispielsweise zwei nicht mit der Zielvariablen korrelierte Attribute nur in Kombination zur Vorhersage beitragen, stößt die einfache logistische Regression an ihre Grenzen.[88] Entscheidungsbäume können diese nichtlinearen Interaktionen ohne Schwierigkeiten lernen und sind einsehbar bzw. intuitiv verständlich. Dementsprechend finden sie in vielen Bereichen Anwendung, in denen Transparenz von durch Daten trainierten Modellen notwendig oder vorteilhaft ist. Entscheidungsbäume sind jedoch ebenfalls nicht für alle Problem- und Datenkonstellationen geeignet. Beispielsweise tendieren sie dazu die Trainingsdaten so gut abzubilden, dass sie auf ungesehenen Daten deutlich ungenauer sind (sogenanntes „overfitting“). Außerdem sind sie bei Regressionsproblemen (Vorhersage von Zahlenwerten statt binären Entscheidungen) anderen Modelltypen unterlegen.

4.4.2.1.3 Zwischenfazit

Im Hinblick auf die Machbarkeit von an den trainierten Modellen direkt ansetzenden Kontrollmechanismen kommen wir somit zu folgenden Ergebnissen:

- Nicht alle durch Daten trainierten Modelle sind von Natur aus transparent im Sinne der Feststellbarkeit und Abschätzbarkeit des Einflusses eines Merkmals auf die Vorhersage.

[85] Calem/Longhofer 2002.
[86] Gängige Lernalgorithmen sind beispielsweise C4.5 (Quinlan 1993) oder CART (Breiman et al. 2017).
[87] Es existieren auch Methoden die mehrere Bäume zu komplexeren Modellen kombinieren (sog. *Random Forests*). In diesem Fall ist die Erklärbarkeit sehr stark eingeschränkt da die einzelnen Bäume zwar einsehbar sind, die Kombinationsmethode aber berücksichtigt werden muss, zumal diese nicht zwangsweise zum Training von intuitiv verständlichen Einzelbäumen führt (siehe Kapitel 4.1.3).
[88] Wie in Abschnitt 4.1 beschrieben existieren zur Modellierung nichtlinearer Abhängigkeiten hierzu sog. „Kernel Tricks/Projections” mit der man die logistische Regression entsprechend erweitern kann. Diese vergrößern die Anzahl der Attribute jedoch erheblich, was wiederum die Interpretation des Modells erschwert.

- Das in der Verbraucherscoringpraxis weit verbreitete Verfahren der logistischen Regression ist ein vergleichsweise transparentes statistisches Modell, dessen gelernte Parameter untersuchbar sind, womit es sich gut zur Prüfung eignet. Ein weiteres vergleichsweise transparentes Modell sind Entscheidungsbäume.
- Soweit transparente Architekturen verwendet werden, ist eine Kontrolle der Modelle selbst prinzipiell möglich. Diese können je nach Problem und Daten jedoch komplexeren, weniger transparenten Modellen in ihrer Genauigkeit und Kapazität unterlegen sein. Bei Verwendung von komplexeren Modellen müssen entsprechend zusätzliche bzw. andersartige Indizien angeführt werden, um die Konformität des ADM-Systems mit den rechtlichen Anforderungen zu demonstrieren (siehe nächster Kapitel).

4.4.2.2 Analyse eines Blackbox-Modells mit Daten

Ein Modell, das einem ADM-System zugrunde liegt, besteht im Wesentlichen aus einem Algorithmus zur Vorhersage und trainierten Parametern. Die Wahl des Modells und die Trainingskonfiguration (auch oft „Hyperparameter“ genannt) werden dabei von Entwicklern und Data Scientists zu Anfang festgelegt und können auch im Nachhinein ohne einen erneuten Trainingsvorgang nicht mehr verändert werden. So wird zu Beginn entschieden, welcher Algorithmus fürs Maschinelle Lernen verwendet wird, z.B. neuronale Netze, Naive Bayes oder Entscheidungsbäume. Für jeden dieser Algorithmen existieren anschließend zahlreiche *Hyperparameter*, z.B. maximale Tiefe des Entscheidungsbaumes (sogenanntes *Pruning*) oder Gewichtsänderungsrate des Optimierungsalgorithmus für ein neuronales Netzwerk.[89] Unter Verwendung dieser Konfiguration werden dem Algorithmus Trainingsdaten zugespielt, die er zum Training und zur Generierung der eigentlichen Entscheidungsstruktur verwendet. Er trainiert damit das zugrundeliegende Modell. Dies geschieht durch mathematische Optimierung, z.B. Wahrscheinlichkeitsberechnung bei (Naive) Bayes, oder Informationsgewinn (*Information Gain*) von Attributen bei Entscheidungsbäumen etc. Die Optimierung liefert weitere Parameter, anhand derer sich das Modell ergibt. Das Modell ist somit eine Kombination aus einer Ausgangskonfiguration und einer Parametrisierung, die sich durch das Training ergibt.

Durch das Berechnen und Optimieren hinsichtlich der Eingabeparameter können in den Datenbeständen komplexe Muster und Regelmäßigkeiten, aber auch Anomalien und Unregelmäßigkeiten erkannt werden. Mit welcher Genauigkeit dies durchgeführt werden kann, hängt sehr stark von mehreren Umständen ab, unter anderem von dem ausgewählten Algorithmus des Maschinellen Lernens, den Ausgangsparametern und den Trainingsdaten.

Das trainierte Modell ist eine mathematische Repräsentation der Entscheidungsstruktur und das Ergebnis eines komplexen und vielschichtigen Prozesses. In der Regel ist es jedoch ein deterministischer Prozess, der nachvollziehbar und transparent gemacht werden kann. Das trainierte Modell ist aufgrund seiner eindeutigen mathematischen Repräsentation auch keinem Zufallselement unterworfen. Die mathematische Repräsentation im Falle von Entscheidungsbäumen besteht aus logischen und arithmetischen Regeln, die auf den Eingabeparametern ausgeführt werden. Für Algorithmen des Maschinellen Lernens, die auf Wahrscheinlichkeiten beruhen, werden bedingte Wahrscheinlichkeiten und ein Erwartungswert berechnet. Auch bei sehr komplexen Entscheidungsstrukturen, wie sie beim Training von neuronalen Netzen entstehen können, liegt eine mathematische

[89] Der Begriff „Hyperparameter” wird hier verwendet, da durch sie bestimmt wird, wie das Modell die eigentlichen Modellparameter lernt.

Berechnungsvorschrift zugrunde, die darüber entscheidet, welche Entscheidung auf Basis eines gegebenen Inputs zu treffen ist. Diese kann jedoch sehr komplex und nach menschlichen Maßstäben nicht mehr nachvollziehbar bzw. interpretierbar sein. Ohne zusätzliche Anstrengungen ist es nicht ohne weiteres möglich, die komplexen Entscheidungsstrukturen von vielschichtigen neuronalen Netzen zu verstehen und zu erklären. Die Entscheidungsstrukturen, die als hochdimensionale Matrizen die Gewichte und Assoziationen zwischen den Neuronen und Schichten darstellen, sind zur Interpretation durch einen Menschen bei Netzen nichttrivialer Komplexität im Regelfall nicht geeignet und kommen als Erklärungskomponente dadurch nicht in Frage. Die Sichtbarmachung des Vorhersageverhaltens großer neuronaler Netzwerke ist ein Gebiet aktiver Forschung, jedoch nach unserer Einschätzung noch nicht weit genug entwickelt, um regulativ aufgegriffen zu werden.

Als eine Ergänzung zur Analyse des Gesamtprozesses zur Erstellung eines Systems, das selbstständig Entscheidungen treffen oder vorbereiten kann, ist es jedoch notwendig, auf der Ebene des trainierten Modells die sich ergebende Entscheidungsstruktur – zumindest anteilig – zu berücksichtigen. Die trainierten Modelle können sich sehr stark voneinander unterscheiden.

Es existiert eine Vielzahl weiterer ML-Algorithmen und -Modelle, denen zum Teil gänzlich verschiedene mathematische Methoden zur Repräsentation der Entscheidungsstrukturen zugrunde liegen. Diese können auf Basis von Wahrscheinlichkeiten (z.B. Naive Bayes), Partitionierung des Attributraumes (z.B. Support Vector Machines, Random Forests) oder komplexen neuronalen Netzen funktionieren. Obwohl die Entscheidungsstrukturen vor allem automatisch generiert werden und unter Verwendung von Parametern und Trainingsdaten entstehen, lassen sich die Beziehungen zwischen den Attributen und der Einfluss auf die Entscheidung mathematisch darstellen, wenn auch oft nicht mit einer der logistischen Regression oder Entscheidungsbäumen vergleichbar leichten Interpretierbarkeit.

Vor dem Hintergrund der Detektion von Diskriminierung in ADM wäre es jedoch wünschenswert, die Auswirkungen von Attributen und deren Einfluss auf die Vorhersage zu verstehen. Eine große Herausforderung ist, dass diese Zusammenhänge oftmals nichtlinear sind, von volatilen Lernprozessen abhängen können und der Einfluss einer bestimmten Variablen unter Umständen nicht ohne größeren Aufwand in einem Modell lokalisiert werden kann. Neuronale Netze beispielsweise lernen nichtlineare Interaktionen statistisch durch Zwischenelemente (sogenannte *hidden nodes*) mit, ohne dass diese Zusammenhänge vor der Optimierung klar individuellen Netzwerkelementen zugeordnet wurden. Ein trainiertes neuronales Netz kann auf den Einfluss einer Eingabevariablen überprüft werden, jedoch ist dies je nach Komplexität des Netzwerkes mit erheblichem Aufwand verbunden. Dies kann dazu führen, dass Eingabe-Ausgabe-Beziehungen, obwohl sie mathematisch eindeutig dargestellt werden können, sich einer Interpretation durch den Menschen quasi gänzlich entziehen und entsprechende Modelle als Blackbox behandelt werden müssen.

Es existiert eine Reihe von Techniken aus der Forschung, um solche intransparenten Modelle einsehbar zu machen, sowohl modellspezifisch[90] als auch modellunabhängig. Beispielsweise kann ein beliebiges Blackbox-Modell durch ein Whitebox-Modell „simuliert“ werden (also ein „Modell eines Modells“[91]). Zunächst trainiert man ein intransparentes Modell und benutzt es anschließend, um einen Datensatz vorherzusagen. Die Vorhersagen werden dann als neue Zielwerte eingesetzt (inklusive der Fehler) und ein transparentes

[90] Siehe z.B. für neuronale Netze Andrews et al. 1995.
[91] Siehe z.B. Adler et al. 2018.

Modell (z.B. ein Entscheidungsbaum) wird auf diesen modifizierten Daten trainiert. Im Idealfall schafft das transparente Modell eine nahezu perfekte Emulation des Blackbox-Modells und erlaubt so Einsicht in den Entscheidungsprozess. Diese Methode basiert allerdings unter anderem auf der Annahme, dass beide Modelle eine vergleichbare Lernkapazität haben, was nicht zwangsläufig der Fall ist. Auch müssen sich das Vorhersageproblem und der Datensatz prinzipiell dafür eignen.

4.4.2.2.1 Analyse der Entscheidung auf Instanzebene

Die dritte und abschließende Ebene, auf der ADM analysiert werden kann, um ein vollumfängliches Bild zu bekommen, ist die Instanzebene. Auf dieser sind die Entscheidungen von ADM inklusive des trainierten Modells in Bezug auf einen konkreten Datensatz (also eine Menge von Einzelinstanzen von Entscheidungen) verortet. Während sich die Analyse des ADM-Prozesses vor allem darauf beschränkt, alle zur Entwicklung und Nutzung eines ADM-Systems notwendigen Schritte transparent und nachvollziehbar zu machen, werden auf der Instanzebene Transparenzkriterien zur Nachvollziehung einer einzigen, für sich abgeschlossenen Entscheidung gesammelt und analysiert. Analog kann diese Ebene auch von der Modellebene abgegrenzt werden, da dort die datenunabhängige Analyse der Modellstruktur selbst im Vordergrund steht.

Um den Nachweis der Diskriminierung zu erbringen bzw. zu widerlegen, ist es nicht zwingend notwendig, das trainierte Modell in all seinen Details zu verstehen bzw. vollständig transparent zu machen. Wie im vorangehenden Kapitel erläutert, ist dies oftmals auch nicht trivial bzw. nur durch sehr großen Aufwand zu bewerkstelligen. Es wäre daher sehr erstrebenswert, ADM-Systeme derart zu konstruieren, dass sie für einzelne Entscheidungen neben der Vorhersage auch Informationen dazu liefern, welche Attributwerte der Entscheidungsinstanz (oder deren Kombinationen) für die Entscheidung relevant sind und wie sie durch Modellparameter gewichtet werden. So kann eine interpretierende Person Indizien dafür bekommen, welche Attribute ausschlaggebend für eine Entscheidung waren und welche nicht. Vor dem Hintergrund der Vermeidung von Ungleichbehandlung ist eine solche Funktionalität sehr nützlich. Wie bereits erklärt, erfreut sich die logistische Regression gerade deswegen großer Beliebtheit, weil die intuitive mathematische Struktur einer solchen Sichtung der „Entscheidungsgründe“ sehr nahe kommt.

Im Rahmen der noch sehr jungen Disziplin der sogenannten *Explainable Artificial Intelligence (XAI)* werden Methoden und Ansätze entwickelt, die solche Aussagen über Entscheidungen eines Systems zulassen bzw. ermöglichen. Idealerweise müssen dafür die zugrundeliegenden trainierten Modelle nicht offengelegt werden, sondern die Verfahren behandeln das ADM-System als eine Blackbox und versuchen über intelligente Abfrage- und Testmechanismen die möglicherweise komplexe innere Struktur zu approximieren. Ein solches Verfahren, welches 2016 vorgestellt wurde, heißt „*LIME – Local Interpretable Model-agnostic Explanations*“[92]. Vereinfacht ausgedrückt wird die Entscheidungsstruktur, also das trainierte Modell, im Bereich einer einzelnen Entscheidung durch eine einfache lineare Funktion angenähert. Diese Approximation wird dadurch generiert, dass man für einen bestimmten vorliegenden Datensatz (Instanz) einzelne Attribute ein- und ausblendet bzw. permutiert und das Ergebnis der Entscheidung bzw. die Änderungen der Entscheidung aufzeichnet. Aus dieser Abbildung zwischen Eingabedaten und Entscheidung wird die innere Struktur punktuell angenähert, jedoch nicht vollumfänglich erfasst. Diese Annäherung und die darauf basierende Interpretation dürfen deshalb nur mit Vorsicht erfolgen. Es handelt sich um eine lokale Approximation, weil der Ausgangspunkt ein bestehender Datenpunkt ist.

[92] Ribeiro et al. 2017.

Dieses Verfahren ist ein möglicher Startpunkt, um Indizien der Entscheidungsstrukturen für komplexe trainierte Modelle zu bekommen. Mit zunehmender gesellschaftlicher Relevanz müssen noch mehr solcher Verfahren entwickelt und implementiert werden. Die zusätzliche Transparenz, die man durch solche Verfahren bekommt, stellen weitere Indizien und Methoden dar, die auf eine Diskriminierung in ADM hinweisen können.

4.4.2.2.2 Illustration am Beispiel Kreditscoring

Für den komplexen Bereich des Kreditscorings ist die Differenzierung insoweit hilfreich und notwendig, als die zugrundeliegenden Entscheidungsstrukturen transparent gemacht werden. Darüber hinaus wird für Verfahren, die sich für eine Codeanalyse nicht eignen, der Lebenszyklus, also das Erstellen, Trainieren und Anwenden, dargestellt.

Analyse des Prozesses

Das ausgewählte Verfahren für Kreditscoring muss benannt werden. Anforderungsdokumente wie Pflichten- und Lastenhefte beschreiben die grundlegenden Funktionalitäten und damit strategische Entscheidungen, die dem Einsatz des ADM zugrunde liegen. Auf der Prozessebene lassen sich auch die Datensätze identifizieren, die zum Training des Verfahrens verwendet wurden. Damit lassen sich möglicherweise auch Biases erkennen, die in den Daten bereits angelegt sind (z.B. statistische Tests). Die Vorverarbeitung der Daten, wie beispielsweise das Umwandeln von metrischen Skalen in Ordinalskalen und das Profiling, kann offengelegt werden und indiziell für eine Diskriminierung sein. Die Analyse des Prozesses, in dem das ADM-System erstellt wird, ist also in jedem Fall und unabhängig von dem verwendeten Algorithmus möglich.

Analyse des Modells

Die akquirierten Daten werden verwendet, um ein Modell zu trainieren, das sich je nach verwendetem Modell einer tiefergehenden und vom Menschen einfach zu interpretierenden Analyse entzieht. Wird nachweislich ein Verfahren eingesetzt, das die zur Verfügung stehenden Attribute in einen komplexen Datenraum transformiert (z.B. SVM oder neuronales Netz), so wird es schwierig die Entscheidungsstrukturen sowie die Gewichte zu analysieren. Werden hingegen Entscheidungsbäume oder andere regelbasierte Verfahren verwendet, so ist es durchaus möglich, dass eine Analyse der Modelle durch den Menschen aufschlussreich ist und für Diskriminierung problematische (Teil-)Entscheidungen entdeckt werden können.

Analyse der Entscheidung auf Instanzebene

Auch ohne möglicherweise schützenswerte Geschäftsgeheimnisse über die verwendeten ADM offenzulegen, existieren Verfahren, z.B. *(Metamorphic) Testing*, die es erlauben eine Einzelentscheidung des ADM hinsichtlich Diskriminierung zu beleuchten. Ein Vorteil von ADM ist unter anderem, dass die Entscheidungsfindung kostengünstig ist und keine manuelle Interaktion erfordert. Für eine algorithmische Entscheidung, bei der davon ausgegangen werden kann, dass z.B. anhand des Geschlechts diskriminiert wurde, kann man das Geschlecht verändern (ceteris paribus), sodass man untersuchen kann, ob dies zu einem anderen Ergebnis führt. Hierzu muss man keine weitere Information über die zum Training verwendeten Daten oder das Modell offenlegen. Von einem technischen Standpunkt ist eine wohldefinierte Schnittstelle erforderlich, die solche Abfragen zulässt. Diese Schnittstelle muss nicht notwendigerweise der Öffentlichkeit zugänglich sein, sondern kann auch nur von Aufsichtsbehörden und zertifizierten Stellen verwendet werden.

Für den vorliegenden Untersuchungsgegenstand der Diskriminierung durch Algorithmen im Rahmen von Bonitätsprüfung (Kreditscoring), und überhaupt bei verbraucherrelevanten automatisierten Entscheidungen, sind die Analyse und die Interpretierbarkeit der Entscheidung auf Instanzebene (sowohl anhand eines einzelnen Datensatzes als auch durch Mengen von Testdaten) die zielführendste und vermutlich aussagekräftigste Methode. Dahingehend lautet die Empfehlung, den Fokus auf die Analyse dieser Ebene zu setzen. Für die umfassende Nachvollziehbarkeit des Verhaltens eines ADM-Systems dürfen die beiden anderen Ebenen jedoch nicht vernachlässigt werden.

4.4.2.2.3 Zwischenfazit

Die Analyse von ADM-Systemen muss auf drei Ebenen geschehen, die ineinandergreifen und erheblich zum Verhalten eines ADM-Systems beitragen. Daraus folgt:

- Jeder Schritt bei der Erstellung des ADM-Systems trägt zu dessen Gesamtverhalten bei, muss also dokumentiert bzw. beschrieben werden, um das Gesamtverhalten nachvollziehen zu können.
- Probleme, die bereits bei der Datenerhebung auftreten (z.B. Biases oder unvollständige Daten), können beim Training von ML-Verfahren zu unerwünschtem Verhalten (Fehlern) führen.
- Nicht jedes ML-Verfahren ist zur Erklärung und zur Interpretation durch den Menschen geeignet.
- Auch wenn das trainierte Modell nicht mehr erklärt werden kann, gibt es Verfahren, um Einzelentscheidungen nachzuvollziehen.
- Techniken zur „Sichtbarmachung“ von intransparenten Modellen existieren, sind indes ein Feld aktiver Forschung und gegenwärtig nicht reif zur verpflichtenden Anwendung.

4.4.3 Testen von ADM-Software

4.4.3.1 Grundlagen des Testens von Software

Innerhalb der Disziplin *System und Software Engineering,* also des übergeordneten wissenschaftlichen Feldes, das sich mit dem Design, dem Entwurf und der Implementierung von Softwaresystemen beschäftigt, kennt der Lebenszyklus von Software, und damit von Systemen, die für ADM eingesetzt werden, mehrere Phasen. Diese Phasen lassen sich Kategorien zuordnen, wobei mindestens eine davon der Qualitätssicherung und -sicherstellung dient. Unterschiedliche Verfahren zur Qualitätssicherung von Software, die in den letzten Jahrzehnten entwickelt wurden, sind mittlerweile etabliert und zur Grundlage modernen Software Engineerings geworden. Ein Standardverfahren dabei ist das sogenannte Testen von Software. Dabei handelt es sich um eine Qualitätssicherungsmaßnahme, ähnlich zur Auditierung, bei der ein Softwaresystem eine vorab definierte Eingabe erhält und daraus eine Ausgabe generiert. Die Ausgabe wird anschließend nicht mit einem notwendigerweise vorab definierten Wert verglichen, sondern z.B. mit den Ergebnissen aus anderen Tests, die zur Laufzeit durchgeführt werden.

Gerade beim Testen komplexer und großer Softwaresysteme unterscheidet man Testverfahren, die verschiedene Aspekte und Bereiche eines Systems überprüfen. Man differenziert zwischen mindestens drei Klassen von Testverfahren:

- **Unit Test:** Beim Unit Test wird eine spezifische Menge von Softwarecode, üblicherweise auf der Ebene von Funktionen, getestet. Zu einer Funktion werden in

der Regel mehrere Testfälle geschrieben, um die Funktionalität hinsichtlich Grenz- und Sonderfällen zu beschreiben.

- **Integrationstest:** Bei dem Integrationstest wird das Zusammenspiel von Modulen und Funktionen, die im Produktivsystem zusammenarbeiten, getestet. Die korrekte Funktion auf der Modulebene wurde bereits überprüft und die korrekte Funktion auf einer höheren Ebene, z.B. Gruppe, wird beim Integrationstest sichergestellt.

- **Systemtest:** Bei dem Systemtest wird ein Gesamtsystem mit all seinen Komponenten und Funktionen überprüft. Die Überprüfung komplexer und abstrakter Systemeigenschaften sowie der Interaktion mit anderen Systemen steht hier im Mittelpunkt.

Beim Testen steht immer die Überprüfung funktionaler Eigenschaften eines (Sub-)Systems im Vordergrund. Diese können z.B. sein: eine bestimmte Ausgabe für eine wohldefinierte Eingabe oder eine maximale Laufzeit für die Berechnung eines wohldefinierten Ablaufs etc. Üblicherweise vergleicht man das Ergebnis eines Testfalls mit einem vorab definierten Ergebnis und entscheidet auf Basis dieses Vergleichs, ob ein Test erfolgreich oder nicht erfolgreich war.

Im Bereich der Diskriminierung durch ADM könnte dies z.B. derart erfolgen, dass man dem ADM einen fiktiven Datensatz als Eingabe zur Verfügung stellt und das Ergebnis analysiert. Ein einfaches Testverfahren könnte dann innerhalb des Datensatzes eine Änderung an einem Datumswert vornehmen, z.B. dem Geschlecht, und diesen geänderten Datensatz dem ADM erneut zur Verfügung stellen. Das Ergebnis kann nun mit dem vorherigen Ergebnis verglichen werden, um die Abweichung dahingehend zu bewerten, ob sie indiziell für eine mögliche Diskriminierung ist.

Das Attraktive an Tests ist, dass sie nicht notwendigerweise von Menschen durchgeführt werden müssen, sondern sehr gut automatisierbar sind. Das bedeutet, dass sich Experten die Testfälle überlegen und wiederum als Computerprogramme verfassen. Dies hat den großen Vorteil, dass sie sehr günstig angewendet und durchgeführt werden können.

Nachfolgend soll noch weiterführend auf die Herausforderungen des Testens von ADM eingegangen werden sowie auf moderne Methoden, die für das Überprüfen der Funktionsweise von ADM hilfreich sein können.

4.4.3.2 Herausforderungen beim Testen von ADM

Beim Testen herkömmlicher Softwareprogramme wird ein Datensatz definiert und erzeugt, der dem zu testenden Programm zur Verfügung gestellt wird. Auf Basis dieses Datensatzes wird eine Ausgabe (Ist) erzeugt, die mit einer vorab festgelegten Ausgabe (Soll) verglichen wird. Im Erfolgsfall stimmen beide Ausgabewerte (Soll und Ist) überein; der Test war erfolgreich. Weicht der Soll-Wert vom Ist-Wert ab, so schlägt der Test fehl und das Programm verhält sich nicht wie vorab erwartet.

Dieses Grundprinzip lässt sich auch auf das Testen von ADM übertragen. Wie oben skizziert, kann ein fiktiver Datensatz erstellt werden, der von dem ADM klassifiziert wird. Geringe Permutationen an sensiblen Attributen (Features), z.B. Geschlecht, ethnische Herkunft oder Alter (siehe § 1 AGG), können ohne weiteres durchgeführt und die Veränderungen in den Entscheidungen beobachtet werden. Die Bewertung, ob ein ADM diskriminiert, ist jedoch nicht trivial. Insbesondere deshalb, weil die Anzahl der

Permutationsmöglichkeiten sehr groß werden kann und die entfernte/permutierte Information auch über Proxyvariablen im Modell verbleiben kann.

Beispiel: Ein einfaches ADM-System trifft Entscheidungen und bekommt drei verschiedene Attribute dafür zur Verfügung gestellt:

1. Alter in Jahren: eine ganze Zahl zwischen 0 und 120 Jahren
2. Geschlecht aus einer Auswahl: „männlich", „weiblich", „ohne Angabe"
3. Migrationshintergrund aus einer Auswahl: „ja" oder „nein"
4. Jahreseinkommen: eine ganze Zahl zwischen 0 und 1000 (in tausend Euro)

Mit diesem einfachen Beispiel lassen sich bereits 720 000 (= 120 x 3 x 2 x 1000) verschiedene Datensätze produzieren, die dem System möglicherweise zur Entscheidungsfindung vorgelegt werden. Nicht selten werden jedoch metrische Attribute, z.B. das Alter, auf Kategorien reduziert, etwa „minderjährig" und „volljährig". Damit verringert sich die Menge der möglichen Datensätze bereits enorm. Entscheidend ist jedoch, dass die Menge nichtlinear mit den zur Verfügung stehenden Attributen ansteigt. Im Zeitalter der Digitalisierung, in der E-Commerce-Herstellern unzählige Attribute von Benutzern zur Verfügung stehen, spielt dies eine große Rolle. Ein öffentlich verfügbarer Trainingsdatensatz zum Testen von unterschiedlichen Verfahren im Bereich Kreditscoring beinhaltet 14 unterschiedliche Attribute mit vorwiegend metrischer Ausprägung – damit ist die Menge der möglichen Datensätze sehr groß. Falls keine Obergrenze für ein metrisches Attribut, z.B. das Gehalt, festgelegt ist, ist sie sogar unendlich. Der Datensatz ist unter folgender URL zugänglich (28. Mai 2018): https://github.com/gastonstat/CreditScoring.

Eine weitere Herausforderung, die sich beim Testen von ADM stellt, ist das sogenannte Orakel-Problem. In der Praxis ist es oftmals nicht möglich, die erforderlichen Testdaten zu erstellen bzw. zu einem Testdatensatz die tatsächlich richtige Ausgabe festzulegen. Ein Orakel dient zur Veranschaulichung für ein Konzept, das allwissend ist und die richtige Ausgabe für jede Eingabe kennt. In der Praxis kann es ein solches jedoch aus verschiedenen Gründen nicht geben[93]. Üblicherweise legen Menschen die Ausgabe (Soll-Wert) fest, was für sehr große Datenmengen durchaus schwierig werden kann. Des Weiteren kann nicht sichergestellt werden, dass Menschen für Datenmengen mit sehr vielen unterschiedlichen Parametern den richtigen Ausgabewert festlegen können. Insofern ist das Erstellen von Testdatenmengen eine größere Herausforderung, als man zunächst vermuten möchte.

Beim Erstellen von Testdatenmengen sind verschiedene Aspekte zu berücksichtigen. Zum einen ist die Grundannahme Maschinellen Lernens, dass Trainings- und Testdaten zwar nicht identisch sind, jedoch auf der gleichen bzw. ähnlichen Verteilung beruhen, die die Population repräsentativ abbilden soll. Ist dies nicht der Fall, sind die Evaluierungsmetriken unter Umständen nicht aussagekräftig. Es kann jedoch unter bestimmten Bedingungen sachdienlich sein, die Zusammensetzung der Testdaten bestimmten weiteren Beschränkungen zu unterwerfen. Beispielsweise kann es für einen intuitiv interpretierbaren Test eines Klassifikationsmodells vorzuziehen sein, Testdaten je zur Hälfte aus positiven und negativen Instanzen zu bilden (z.B. gleich viele kreditwürdige und nicht kreditwürdige Männer und Frauen), obwohl in den Trainingsdaten weniger positive als negative Instanzen

[93] Barr et al. 2014.

vorhanden sind. In jedem Fall jedoch müssen Kriterien zur Abdeckung (*Coverage*) der Phänomene in den Daten eingehalten werden. Um ein ADM effektiv zu testen, muss die Datenmenge alle möglichen unterschiedlichen Kombinationen abdecken. Für einfache Entscheidungsverfahren, z.B. regelbasierte Verfahren oder Entscheidungsbäume, kann dies bereits ausreichend sein, um zu erkennen, ob ein Attribut eine Entscheidung unverhältnismäßig stark beeinflusst. Durch komplexe Verfahren, die auf Wahrscheinlichkeiten oder hochdimensionalen Operationen beruhen, ist ein Rückschluss auf die zugrundeliegende Entscheidungsstruktur nur sehr schwer möglich. Ein einfacher Rückschluss auf Basis von wenigen Datenpunkten anhand einzelner Kriterien mag naheliegend erscheinen, wird aber unter Umständen der Komplexität des ADM nicht gerecht.

Darüber hinaus ist es notwendig die Phasen des Lebenszyklus eines ADM-Systems zu unterscheiden. Die zugrundeliegenden Algorithmen bzw. ihre Parameter können sich, je nach eingesetztem Verfahren, weiterentwickeln. Selbstlernende Systeme verändern ihre Entscheidungsstrukturen über die Zeit. Dies muss berücksichtigt werden. Diese Veränderung erfolgt in der Regel aber nicht zufällig, obwohl auch solche Systeme existieren und eingesetzt werden könnten[94]. Die Veränderungen lassen sich jedoch immer nachvollziehen. Das bedeutet, dass es prinzipiell denkbar ist, die Parameter und Datensätze, die zur Weiterentwicklung (Training) verwendet wurden, zu speichern. Damit lässt sich der aktuelle Zustand eines ADM-Systems jederzeit reproduzieren. Dieses Vorhalten der Daten kann sehr datenintensiv sein und damit eine Herausforderung, die sich beim Testen von ADM-Systemen notwendigerweise ergibt.

4.4.3.3 Methoden zum Testen von ADM: Metamorphic Testing (MT)

Das Testen von Software kann mitunter sehr aufwändig sein. Dies gilt insbesondere dann, wenn die möglichen Eingabemengen sehr groß sind. Diese können zwar maschinell erstellt werden, jedoch muss für jeden Datensatz auch eine erwartete Ausgabe definiert werden. Dieses sogenannte Orakel-Problem wurde beim Testen von Software schon sehr früh erkannt und kann insbesondere zum Problem werden, wenn ADM-Systeme überprüft werden. Eine Möglichkeit, wie man sich diesem Problem stellen kann, ist das sogenannte *Metamorphic Testing* (MT).

Das Grundprinzip von MT ist den herkömmlichen Testverfahren sehr ähnlich. Der Hauptunterschied liegt jedoch darin, dass nicht mehr einzelne Datensätze verarbeitet und das Ergebnis (Ist) mit einem vorab definierten Ergebnis (Soll) verglichen wird, sondern Beziehungen (sogenannte Relationen) festgelegt werden, deren Validität einfacher überprüft werden kann. Diese Relationen definieren die Beziehung zwischen den Eingabedaten und der Ausgabe, in diesem Kontext der Entscheidung des ADM.

Am Beispiel der Bonitätsprüfung von oben könnte dies wie folgt aussehen:

Beispiel: Das oben genannte ADM-System bekommt einen fiktiven Datensatz zur Bonitätsprüfung:

1. Alter: 50 Jahre
2. Geschlecht: „männlich“
3. Migrationshintergrund: „nein“

[94] Sogenannte Evolutionäre Algorithmen.

Das System liefert einen Wert (*Score*) für die Wahrscheinlichkeit des Kreditausfalls von 20 (auf einer Skala von 0 bis 100). Herkömmliche Testverfahren würden diesen Wert mit einem vorab definierten Wert vergleichen. Dies ist problematisch, weil der Wert kaum von Experten vorab festgelegt werden kann.

MT-Verfahren könnten die Eingabedaten variieren und folgende Relation festhalten:

`(50, männlich, nein) → 20`

`(50, weiblich, nein) → XX`

Für den zweiten Datensatz, der nur geringfügig angepasst wird (geändertes Attribut unterstrichen), würden nun erneut der Score-Wert und der Unterschied zum ursprünglichen Wert berechnet werden. Damit werden nicht mehr die absoluten Werte berücksichtigt, die vom ADM-System berechnet werden, sondern die Veränderung, die sich durch die Permutation ergibt. Dieser Wert kann indiziell für eine problematische Differenzierung, z.B. Diskriminierung, sein.

Da MT den Fokus auf die Relation, also auf das Verhältnis von Ein- und Ausgabe, legt und die Änderungen der Ausgabe (Entscheidung) in den Vordergrund rückt, ist ein wohl-definierter Datensatz, wie er bei herkömmlichen Testverfahren notwendig ist, hinfällig. Damit ist MT ist der Lage, das Orakel-Problem zu umgehen.

4.4.3.4 MT und Testen von ADM

Testen ist eine Möglichkeit, die Transparenz einer Entscheidung bzw. der gesamten Entscheidungsstruktur eines ADM-Systems zu erhöhen. Unterschiedliche Ebenen (Prozess, Modell und Entscheidung) erfordern dabei eine differenzierte Betrachtung. MT kann Indizien für eine möglicherweise Differenzierung anhand unzulässiger personenbezogener Merkmale und damit für eine Diskriminierung liefern.

Eine Herausforderung, die sich allerdings auch durch das Testen nicht oder nur sehr schwer in Griff bekommen lässt, ist, dass alle Eingabedaten durch das System festgehalten werden müssen. Insbesondere dann, wenn ein ADM-System auf Daten operiert, die es wiederum von anderen System bezieht, z.B. die Verfügbarkeit von Gütern auf anderen Onlineshops, oder auf externe physikalische Größen, z.B. Uhrzeit und Datum, zurückgreift, wird das Testen sehr schnell zu einer komplexen Herausforderung.

Durch Testen kann man Indizien sammeln, die die Eingabeattribute und deren Beitrag zu einer Entscheidung ein Stück weit transparenter machen. Für einzelne Entscheidungen kann dies durchaus hilfreich sein. Die Frage nach dem Nachweis einer unzulässigen Diskriminierung ist damit aber noch nicht vollständig beantwortet. Dazu muss der Diskriminierungsbegriff für den Einzelfall (oder eine gesamte Branche) noch stärker operationalisiert werden. Es muss also geklärt sein, welchen Einfluss die Anwesenheit bzw. Abwesenheit eines Attributs maximal haben darf: Ab wann kann man von einer Diskriminierung sprechen und inwieweit dürfen Attribute zur Differenzierung von Merkmalen eine Rolle spielen? Diese Fragen können aus einer Betrachtung der Technologie nicht beantwortet werden. Sie berühren die Diskussion um Fairness im Bereich des Maschinellen Lernens und bedürfen der expliziten Berücksichtigung. Eine umfassendere Diskussion von Fairness in Klassifikation wurde in Kapitel 4.3 durchgeführt; die Verbindung zu juristischen Begriffen der Ungleichbehandlung besprechen wir sodann in Kapitel 5.4.3.

4.4.3.5 Illustration am Beispiel Kreditscoring

Wie bereits erwähnt kann für die Berechnung der Bonität einzelner Personen aus einer sehr großen Anzahl von Algorithmen, die ganz unterschiedlich parametrisiert werden können, ausgewählt werden. Das Testen behandelt diese Algorithmen wie eine Blackbox, das bedeutet, dass die intern verwendeten Algorithmen sowie die zum Training benutzten Daten nicht offengelegt werden müssen.

Das Testen eines Verfahrens zur Bonitätsprüfung könnte wie folgt aussehen:

Beispiel: Ein Unternehmen U verwendet ein ADM-System A zur Bonitätsprüfung. Ein Benutzer muss dazu Angaben über sich selbst machen (z.B. in einem Onlineformular). Ein fiktiver Datensatz zur Bonitätsprüfung könnte wie folgt aussehen:

1. Alter: 35 Jahre
2. Geschlecht: „männlich“
3. Migrationshintergrund: „nein“
4. Jahreseinkommen: 50.000 €

```
(35, männlich, nein, 50 T) → 75
```

A liefert einen Score, der die Wahrscheinlichkeit des Kreditausfalls repräsentiert (z.B. durch die Schufa[95]) von 75 (auf einer Skala von 0 bis 100). Herkömmliche Testverfahren würden diesen Wert mit einem vorab definierten Wert vergleichen. Dies ist problematisch, weil der Wert kaum von Experten vorab festgelegt werden kann (Orakel-Problem).

Um dennoch Indizien für eine mögliche Diskriminierung zu finden, werden fiktive Testdaten erzeugt und die Ergebnisse der automatisierten Entscheidung verglichen.

MT-Verfahren könnten die Eingabedaten variieren und folgende Relation festhalten:

```
I     (50, männlich, nein, 50 T) → 80 → (+ 5)

II    (35, weiblich, nein, 50 T) → 85 → (+ 10)

III   (35, männlich, ja,    50 T) → 70 → (− 5)

IV    (35, männlich, nein, 100 T) → 90 → (+ 15)

V     (20, männlich, ja,   0 T)  → 20 → (− 55)
```

Für die fiktiven Datensätze I bis IV, in denen jeweils nur ein Attribut geändert wurde, würden nun erneut der Score und der Unterschied zu dem ursprünglichen Wert berechnet werden. Für den Datensatz I beispielsweise hat sich der Score durch das Verändern des Attributs Alter von 35 auf 50 Jahre um +5 erhöht. Im Datensatz III hingegen, der den Migrationshintergrund von „nein“ auf „ja“ geändert hat, verringert sich der Score um 5.

Damit werden nicht mehr die absoluten Scores berücksichtigt, die vom ADM-System berechnet werden, sondern die Veränderung, die sich durch die Permutation ergibt. Dieser

[95] Siehe www.bonify.de/schufa-score (letzter Zugriff am 6. April 2018).

Wert kann indiziell für eine problematische Differenzierung, z.B. Diskriminierung, sein. Siehe hier auch die Ausführungen zu quantitativen Fairnessbegriffen in Kapitel 4.3 und das dortige Beispiel.

Die Datensätze für das Testen können sehr schnell, z.B. automatisch, erzeugt werden. Auch das Abfragen des ADM-Systems kann sehr effizient durchgeführt werden, sofern eine entsprechende Programmierschnittstelle (API) von dem Betreiber zur Verfügung gestellt wird. Führt man diese Tests strukturiert durch, so erhält man einen guten Überblick über die Funktionsweise des ADM-Systems, ohne dessen interne Struktur zu kennen. Das System wird ein Stück weit transparent und Entscheidungsverfahren lassen sich möglicherweise rekonstruieren. Testverfahren werden deshalb auch häufig als eine Form des Audits angesehen.

4.4.3.6 Zwischenfazit

Testen ist eine geeignete Methode, um das Verhalten von ADM-Systemen zu beschreiben, ohne deren interne Struktur exakt kennen bzw. offenlegen zu müssen. Für diesen Einsatzzweck zeichnen sich einige Punkte ab:

- Herkömmliche Testverfahren sind nur bedingt zur Detektion von fehlerhaftem Verhalten geeignet, weil es nicht möglich ist das zu erwartende Ergebnis vorab zu definieren (Orakel-Problem).
- Metamorphic Testing ist hier eine geeignete Methode, um das Verhalten auf sich ändernden Input zu beschreiben.
- Testen kann günstig und effizient durchgeführt werden. Die Erstellung eines Testdatensatzes variiert von System zu System und bedarf technischer und fachlicher Expertise.
- Der Testdatensatz kann durch einfache Permutation der Eingabedaten das Verhalten und die Reaktion auf das Verfahren für einzelne Attribute (z.B. Alter oder Geschlecht) beschreiben und offenlegen.

4.4.4 Auditing von im Betrieb befindlichen ADM-Systemen

Die Auditierung von Softwaresystemen, und damit auch Modellen von ADM-Systemen, ist ein Standardverfahren bei der Sicherung von Qualität und Zuverlässigkeit. Bei der Umsetzung und zur Kontrolle des Fair Lending Acts in den USA untersucht das Aufsichtsorgan bestehende Algorithmen und Verfahren hinsichtlich der Diskriminierung. Um Indizien für eine solche zu sammeln, werden statistische Überprüfungen durchgeführt[96]. Die Auditierung der Algorithmen geht noch einmal deutlich weiter, weil sie in bestimmten Fällen die Offenlegung der Algorithmen, der verwendete Daten und der Trainingsmethoden bedeutet, die sensible Geschäftsgeheimnisse sein können. Dennoch ist es naheliegend, Auditierung auf Algorithmen anzuwenden, die im Rahmen der ADM angewandt werden.

Die Übertragbarkeit von Auditierungsverfahren, die in der analogen Welt durchaus üblich sind, wurde bereits nachgewiesen.[97] Nicht ohne weiteres können vorhandene Methoden zur Auditierung von IT-Systemen übertragen werden. Nach Sandvig existieren fünf Strategien, die sich als Auditierungsverfahren für ADM eignen:

[96] Die genaue Praxis bei der Prüfung in den USA wurde beschrieben von Calem/Longhofer 2002.
[97] Sandvig 2014.

1. Code Audit
2. Nichtinvasive Audits
3. Scraping Audits
4. Sock Puppet Audits
5. Crowdsourced Audits

Nachfolgend werden fünf Grundstrategien zur Auditierung im Detail vorgestellt und anschließend wird deren Praktikabilität für ADM diskutiert.

4.4.4.1 Code Audits

Code Audits sind eine naheliegende Methode zur Überprüfung der Funktionsweise eines Algorithmus, auch als *„Algorithm Transparency“* bezeichnet. Im Falle des Verdachts eines fehlerhaften (oder diskriminierenden) Entscheidungsverhaltens wird eine Kopie des Algorithmus erstellt und von einem unparteiischen Dritten, dem Auditor, begutachtet. Der Auditor analysiert den Algorithmus und seine implementierten Entscheidungsstrukturen und kann somit, im Idealfall, das Entscheidungsverhalten erkennen und darüber Auskunft geben, ob ein Fehlverhalten oder ein anderweitiges nicht erwünschtes bzw. nicht wünschenswertes Verhalten, z.B. Diskriminierung, vorliegt oder nicht.

Das Verfahren kann wie folgt illustriert werden (Sandvig 2014):

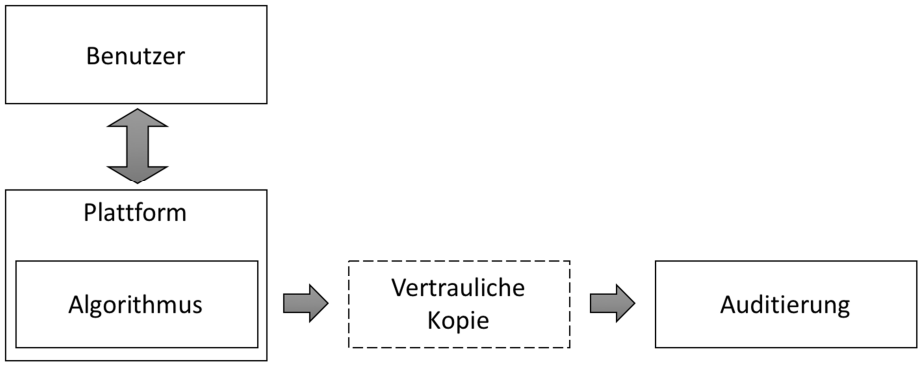


Abbildung 4: Schematische Darstellung des Code-Audit-Verfahrens

In Abbildung 4 ist der Ablauf des Audits in groben Zügen dargestellt: Der Algorithmus, der in einer Plattform zur Entscheidungsfindung über Benutzer und Konsumenten eingesetzt wird, wird als eine kopierte Version an die Auditierungsstelle weitergereicht. Diese Kopie enthält alle Daten und Zustände, die bei der benutzten Version des Algorithmus verwendet werden, und ist somit repräsentativ für den Entscheidungsfindungsprozess. Die Kopie enthält die gesamte Information in einer nichtverschlüsselten Repräsentation und kann somit von der Auditierungsstelle analysiert werden. Wichtig dabei ist es, dass die Kopie vertraulich behandelt wird und ausschließlich privilegierten und vertrauenswürdigen Dritten zugänglich ist. Die Kopie kann wertvolle Daten über Geschäftsvorfälle, Kundeninformationen und Unternehmensstrategien enthalten.

Unter den Nachteilen des Code Audits sind zwei Punkte besonders hervorzuheben:

1. Komplexe Algorithmen erfordern sehr hohen Aufwand bei der Überprüfung
2. Sich verändernde Entscheidungsstrukturen müssen immer wieder überprüft werden

Es gibt kaum verlässliche Literatur, die darüber aufklärt, wie ADM in Unternehmen in den Bereichen Human Resources oder Kreditscoring aktuell eingesetzt wird. Wie jedoch in Kapitel 3.1 und 3.2 erläutert, existieren an der Schnittstelle zwischen ADM und ML sehr komplexe Verfahren, die zwar rational sind (im mathematischen und deterministischen Sinne), für Menschen jedoch nur mehr sehr schwer nachzuvollziehen. Zur Überprüfung hinsichtlich Diskriminierung ist nicht nur spezifisches Expertenwissen, sondern auch ein hoher Ressourcenaufwand notwendig. Darüber hinaus sind konkrete überprüfbare Kriterien erforderlich, anhand derer sich Diskriminierung bestimmen lässt. Die zweite Herausforderung bei der Auditierung von eingesetzten Algorithmen bei ADM ist die hohe Dynamik, mit der sich diese weiterentwickeln. Die Verfahren des Maschinellen Lernens werden kontinuierlich weiterentwickelt. Damit verändern sich nicht nur die internen Parameter des Modells, sondern auch das Entscheidungsverhalten kann sich über die Zeit hinweg verändern. Damit reicht es also nicht aus den Algorithmus einmalig zu analysieren, um sein Verhalten zu bewerten, da sich dieses kontinuierlich verändert.

4.4.4.2 Nichtinvasive Audits

Neben der Möglichkeit der Analyse des konkreten auszuführenden Programmcodes des Algorithmus gibt es die Möglichkeit der sogenannten nichtinvasiven Audits. Diese stellen nur im weitesten Sinne eine Form der Auditierung dar, weil es nicht, wie bei den zuvor diskutierten Verfahren, darum geht, die Funktionsweise des Algorithmus über gesicherte Verfahren zu rekonstruieren, sondern über die Rückmeldung von Benutzern. Benutzer werden dazu befähigt ihr eigenes Verhalten und Nutzen zu beschreiben und in einer Art Fragebogen festzuhalten. Über größere Mengen von Datenbeständen, die die Nutzungsweise der Plattform beschreiben, lassen sich dann Rückschlüsse auf das Verhalten und die (automatisierte) Entscheidungsfindung innerhalb der Plattform ziehen.

Der Ablauf könnte dabei wie folgt dargestellt werden:

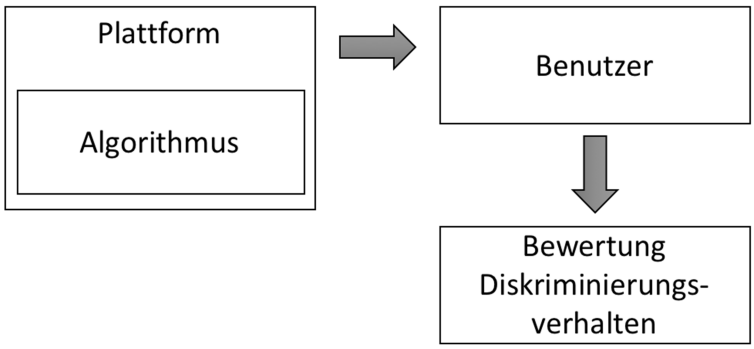


Abbildung 5: Schematische Darstellung nichtinvasiver Auditverfahren

Benutzer interagieren mit einer Plattform, z.B. im Bereich Kreditscoring, und beschreiben die Daten, die sie der Plattform zur Verfügung gestellt haben, und welche Rückmeldung sie daraufhin bekommen haben. Um diese Daten besser verwerten zu können, erfolgt dies idealerweise in Fragebögen oder Ähnlichem. Liegt eine größere Anzahl an Rückmeldungen über das Verhalten und die Entscheidungen einer Kreditscoringplattform vor, so kann man Aussagen über das Verhalten treffen und z.B. über statistische Verfahren Indizien für eine mögliche Diskriminierung finden (zur Diskussion des Fairnessbegriffs siehe Kapitel 4.3).

Dieses Verfahren bedarf natürlich einer repräsentativen Datenmenge, um überhaupt statistisch signifikante Aussagen über eine Plattform machen zu können. Die Erhebung der

Daten (Sampling) stellt eine große Herausforderung dieser Methode dar, da hier auch sensitive (persönliche) Daten erhoben und persistiert werden müssen.

4.4.4.3 Scraping Audits

Bei der Überprüfung des Verhaltens eines Algorithmus existiert die Möglichkeit, sein Verhalten über automatisierte Verfahren zu erfassen. Dabei werden kleine Programme, z.B. Skripte, entwickelt und ausgeführt, die wiederholt Anfragen an einen in einer Plattform eingesetzten Algorithmus senden und die Antworten auswerten. Dieser Ansatz ähnelt dem automatisierten Testen von Software, welches bei Software Engineering im industriellen Kontext sehr etabliert ist. Das Ziel dabei ist es, ein bestimmtes Verhalten der entwickelten Software zu vermeiden bzw. sicherzustellen. Das Skript entwickelt Datenbestände und sendet diese an den Algorithmus, der eine Entscheidung auf diesen (fiktiven) Daten trifft. Unter Verwendung dieser Skripte ist es möglich eine große Bandbreite an möglichen Entscheidungen zu analysieren und zu bewerten.

In der Regel erfolgt die Abfrage nicht über die gleiche Schnittstelle, über die Benutzer ihre Anfragen stellen, sondern über eine separate API. Der Algorithmus wird dabei als eine Blackbox betrachtet, seine internen Entscheidungsstrukturen müssen nicht offengelegt werden und verlassen die Plattform und somit das Unternehmen nicht.

Das Verfahren ist in folgender Abbildung illustriert:

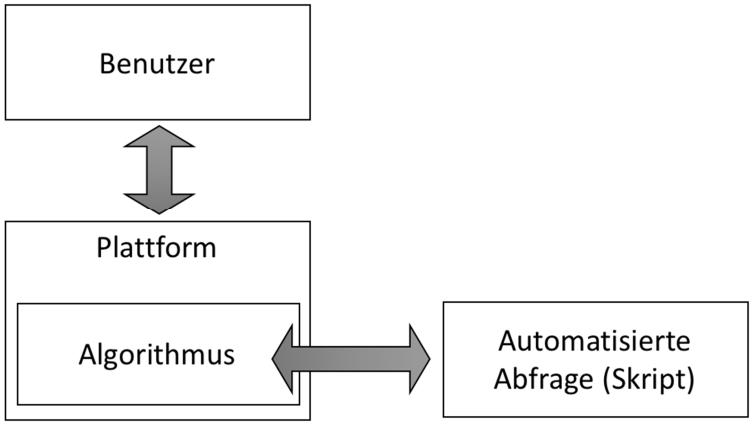


Abbildung 6: Schematische Darstellung von Scraping-Audit-Verfahren

Die Abfrage aus einem Skript heraus erfolgt automatisch und ist in der Lage, eine Rückmeldung zu verschiedenen Variationen von Daten zu erhalten. Diese Daten werden derart generiert, dass eine Aussage über das Entscheidungsverhalten getroffen werden kann. Dies könnte z.B. so aussehen, dass in einem Datensatz nur kleine Änderungen an Attributen vorgenommen werden, deren Unterscheidung indiziell für Diskriminierung ist (z.B. Geschlecht). Verändert der Algorithmus auf Basis dieser kleinen Veränderung der Daten seine Entscheidung, so liegt die Vermutung nahe, dass die interne Entscheidungsstruktur problematische Attribute hinsichtlich Diskriminierung berücksichtigt. Der Vorteil dieses Verfahrens liegt in der automatisierten Anwendung und den im Vergleich zur manuellen Code-Analyse niedrigen Transaktionskosten. Ändert sich der Algorithmus, kann das Skript erneut gestartet und die Analyse angestoßen werden. Anders als beim nichtinvasiven Audit werden auch keine sensiblen Daten von Benutzern benötigt, sondern nur fiktive (generierte) Daten verwendet.

4.4.4.4 Sock Puppet Audits

Das Sock Puppet Audit ist eine Weiterentwicklung des Scraping Audits und insbesondere für die Analyse von Plattformen geeignet, deren Algorithmus zur Entscheidungsfindung sich nicht direkt, z.B. über eine API, ansprechen lässt. Es ist technologisch möglich, Computerprogramme derart zu schreiben, dass sie das Verhalten von Benutzern simulieren und sich gegenüber der Plattform so verhalten, als wären sie menschliche Benutzer. Solche Computerprogramme nennt man üblicherweise „*sock puppets*“. Diese fiktiven Benutzer sind mit Strategien ausgestattet, um sich gegenüber der Plattform mit verschiedenen Attributen darzustellen. Die Menge an mögliche Attributen und deren Ausprägung ist variabel und kann je nach Strategie gesteuert und angepasst werden.

Das Verfahren kann anschaulich wie folgt dargestellt werden:

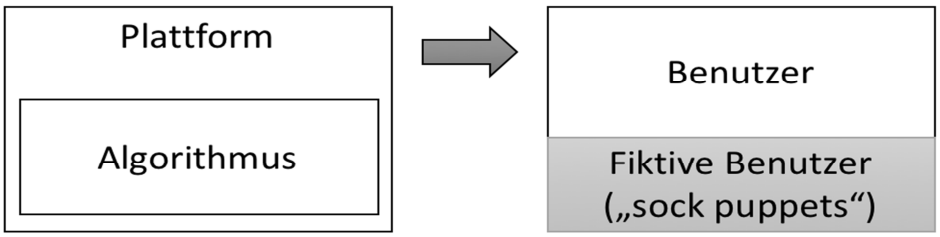


Abbildung 7: Schematische Darstellung von Sock-Puppet-Audit-Verfahren

Auf die Plattform wird von außen über Schnittstellen zugegriffen, die auch menschlichen Benutzern zur Verfügung stehen. Die fiktiven Benutzer *(sock puppets*) senden auf die gleiche technische Art und Weise Anfragen und versuchen dabei die Anfragen derart zu strukturieren und zu gestalten, dass Aussagen über die zugrundeliegenden algorithmischen Entscheidungsstrukturen getroffen werden können. Über die Variation kleinerer Änderungen in den Datensätzen, die automatisch generiert werden, kann das Verhalten der Plattform über einen großen Bereich erfasst und abgeglichen werden. Idealerweise lassen sich dann auch Rückschlüsse darüber ziehen, ob und in welchem Ausmaß sensible Attribute Einfluss auf das ADM haben.

Der Vorteil dieses Verfahrens besteht darin, dass die Kosten zur Durchführung verhältnismäßig gering sind und sich ein sehr breites Spektrum an möglichen Benutzergruppen abbilden lässt.

Analog zum Scraping Audit ist auch das Sock Puppet Audit nicht ohne weiteres auf Internetplattformen anwendbar. Die praktische Hürde dabei ist die technologische Ausgestaltung der Schnittstelle, die eine automatisierte Abfrage durch Skripte oder Computerprogramme aus Sicherheitsgründen oftmals nicht zulässt. Zusätzlich wird in manchen Jurisdiktionen das automatisierte Abfragen als eine illegale Maßnahme zur Analyse und zur Rekonstruktion der Plattform (samt Algorithmus), also von schützenswerten Betriebsgeheimnissen, angesehen.[98]

4.4.4.5 Crowdsourced Audits

Als eine Variante des Sock Puppet Audits kann die Auditierung durch die „Crowd“ angesehen werden. Anstelle von programmierten fiktiven Benutzern bestimmt man eine Menge von „affiliierten“ Benutzern, zu denen man ein besonderes Vertrauensverhältnis

[98] Siehe dazu auch Sandvig 2014, S. 11.

aufbaut. Diese Benutzer bekommen den Auftrag, eine Plattform und den zugrundeliegenden Algorithmus während des Betriebs zu testen. Diese Form des Audits wurde auch als eine Strategie angesehen, um nicht von den Einschränkungen und Verboten betroffen zu sein, die sich bei der automatisierten Auditierung der Plattformen von außen ergeben (siehe Scraping-Audit und Sock-Puppet-Audit).

Die Aufgaben, die dort ein Computerprogramm innehat, nämlich die Abfrage mit unterschiedlichen und repräsentativen Datumswerten, übernehmen wieder menschliche Benutzer.

Eine grafische Darstellung könnte wie folgt aussehen:

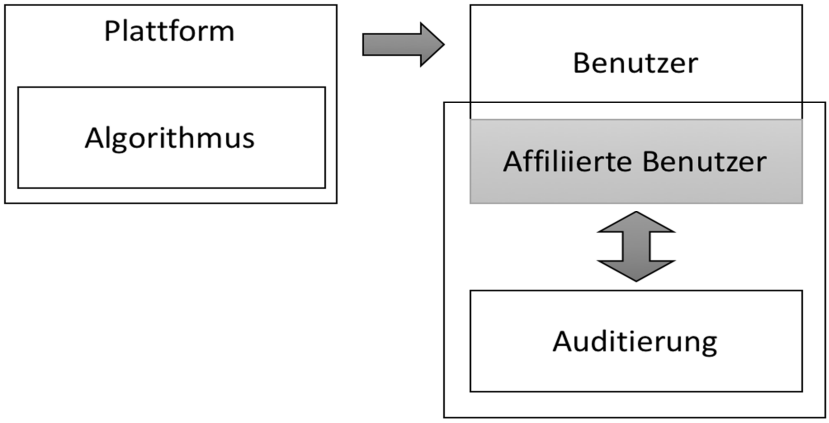


Abbildung 8: Schematische Darstellung von Crowdsourced-Audit-Verfahren

Die Illustration zeigt die Interaktion der Benutzer mit der Plattform, in welche der Algorithmus eingebettet ist. Im Kreis der Benutzer ist eine Menge speziell gekennzeichnet, die im Austausch mit der Auditierungsstelle steht. Diese Stelle orchestriert die affiliierten Benutzer dahingehend, als dass sie diese anleitet und ihnen die unterschiedlichen Parameter mitteilt, anhand derer das ADM bewertet werden soll.

4.4.4.6 Illustration am Beispiel Kreditscoring

Code Audits: Beim Code Audit wird einem Dritten, z.B. einer vertrauenswürdigen Behörde, der gesamte Programmcode zur Verfügung gestellt, der die Algorithmen beinhaltet, die innerhalb das ADM-Systems verwendet werden. Dieser Code umfasst nicht nur die trainierten Algorithmen, sondern auch die Programmcodes, die zur Datenerhebung, Vorverarbeitung und Nachbearbeitung verwendet werden (siehe Kapitel 3.4). Die Einbettung von ADM-Systemen in einen komplexen Prozess macht dies erforderlich – denn ohne Berücksichtigung dieser Elemente lassen sich keine vollumfänglichen Rückschlüsse auf die Entscheidungsstrukturen und deren Zustandekommen ziehen.

Nichtinvasive Audits: Beim nichtinvasiven Audit werden Benutzer eines Kreditscorings über das Systemverhalten befragt und deren Antworten ausgewertet. Benutzer fragen eine Bonitätsprüfung an, ohne besondere Vorkehrungen zu treffen, und bewerten danach, welchen Eindruck das System auf sie macht bzw. ob sie eine Entscheidung (nicht) nachvollziehen können. Dies kann auf Basis von Interviews oder Fragebögen stattfinden.

Scraping Audits: Das Kreditscoringsystem stellt eine Programmierschnittstelle zur Verfügung, die, nicht notwendigerweise öffentlich, zur Verfügung gestellt wird. Diese

Programmierschnittstelle kann mit automatischen Skripten abgefragt werden. Diese Skripte können eine große Anzahl an Abfragen machen und möglicherweise noch weitere Informationen, wie etwa Konfidenz einer Entscheidung, Alternativen und Erklärungen (z.B. Gewichte einzelner Attribute) erhalten. Die Programmierschnittstelle kann möglicherweise noch mehr Informationen anbieten, als normale Endverbraucher bekommen.

Sock Puppet Audit: Beim Sock Puppet Audit werden fiktive Benutzer simuliert (ähnlich dem Scraping Audit), in der Regel durch andere Computerprogramme, die sich dem Kreditscoringsystem gegenüber als Benutzer bzw. Personen ausgeben, als würden sie das System normal benutzen wollen. Hierbei werden unterschiedliche Parameter, z.B. Herkunft, Kaufverhalten oder Browserverlauf, geändert, sodass eine umfangreiche Aussage über das Entscheidungsverhalten des ADM-Systems getroffen werden kann.

Crowdsourced Audit: Analog zum Sock Puppet Audit werden die Daten von unterschiedlichen Benutzern eines Systems (z.B. Schufa) ausgewertet und analysiert. Im Gegensatz zu Sock Puppets willigt eine Gruppe, die Crowd, in die Bereitstellung der eigenen Daten und des Ergebnisses des ADM-Systems in ein Drittsystem ein. In diesem Drittsystem werden die Daten konsolidiert und über die große Datenmenge ausgewertet.

4.4.4.7 Zwischenfazit

Audits sind eine geeignete Maßnahme, um die Qualität von ADM-Systemen zu überprüfen und mögliches Fehlverhalten zu identifizieren. Audits und Testen überlagern sich in mehreren Aspekten sehr stark und ergänzen sich in diversen Bereichen. Für das Auditieren von ADM-Systemen kann Folgendes festgehalten werden:

- Es gibt mehrere Arten, ein ADM-System zu auditieren.
- Die Audits reichen bis zur vollständigen Offenlegung des Codes sowie aller Dokumente und Zwischenschritte, die erfolgt sind, bis das ADM-System zum Einsatz gelangt.
- Die vollständige Transparenz steht möglicherweise mit den Geschäftszielen und Geschäftsgeheimnissen im Konflikt.
- Auditierung kann auch eine strukturierte Abfrage des ADM-Systems beinhalten; hierbei wird sie dem Testen (siehe Kapitel 4.4.3) sehr ähnlich. Dabei muss nicht notwendigerweise der Code offengelegt werden, sondern das System wird als eine Blackbox beschrieben. Auch hier sind Rückschlüsse auf möglicherweise vertrauliche Entscheidungsstrukturen des ADM-Systems möglich.
- Betrachtet man Testverfahren als Kontrollmechanismen für ADM-Systeme, so zeigt sich, dass das *Scraping Audit* und das *Sock Puppet Audit* in weiten Teilen dem Testen entsprechen. Die Verfahren nähern sich an und lassen sich kaum mehr voneinander abgrenzen, was für die Verwendung als Kontrollmechanismus auch nicht notwendig ist.
- Auditierung bedarf einer technischen Expertise sowie der fachlichen, juristischen Expertise, um die Ergebnisse entsprechend aufbereiten und interpretieren zu können.

4.4.5 Auditing von archivierten ADM-Systemen

Da es sich bei ADM-Systemen auch um Softwaresysteme handelt, unterliegen sie in der Regel einer ständigen und fortlaufenden Änderung und Anpassung. ADM-Systeme werden weiterentwickelt und weisen möglicherweise ein neues Entscheidungsverhalten auf. Um eine rückwirkende Kontrolle zu ermöglichen, ist es notwendig die Entscheidungsmodelle zu

archivieren und bei Bedarf wieder einspielen zu können. Eingespielte Modelle aus den Archiven sind der Analyse, dem Audit sowie dem Testing wie oben beschrieben zugänglich.

Das Archivieren von trainierten Modellen ist unter Umständen ressourcenintensiv (insbesondere Speicherplatz), technisch jedoch machbar. Speziell die Verwendung von Software zur Versionskontrolle ist hier geeignet. Die Speicherung von trainierten Modellen erscheint für die Nachvollziehbarkeit und Rekonstruierbarkeit von Entscheidungen sinnvoller als die Speicherung von großen Trainingsdatenbeständen etc. Bei selbstlernenden ADM-Systemen, die ihr Verhalten während des Betriebs weiterentwickeln, reicht es jedoch nicht, die Entscheidungsmodelle „von Zeit zu Zeit“ zu archivieren. Hier bliebe die Möglichkeit, einen „Snapshot“ vor jeder Entscheidung zu archivieren, was in der Praxis nicht realisierbar ist, oder die jeden zum Training verwendeten Datensatz zu speichern. Beide Möglichkeiten sind für die Praxis vermutlich unbefriedigend und – wenn überhaupt – nur durch sehr hohen Aufwand durchführbar.

4.5 Fazit

Im vorliegenden Kapitel haben wir die Grundkonzepte des Maschinellen Lernens sowie die logistische Regression als Beispiel eines praktisch relevanten und einfach prüfbaren statistischen Klassifikationsmodells dargestellt. Aus den Grundlagen des Maschinellen Lernens ergeben sich auch unmittelbar die praktischen Probleme, die unter anderem in Diskriminierung münden können: Unausgewogenheiten in den Daten können sich im Trainingsprozess niederschlagen. Die Nichtverwendung von geschützten Attributen (aufgrund derer eine Diskriminierung also nicht zulässig ist) ist keine universelle Lösung, da einerseits die Gruppenzugehörigkeit mittels Proxyvariablen für das Modell erkennbar bleiben kann und andererseits die Genauigkeit des trainierten Modells reduziert werden kann. Somit ergibt sich ein potenzieller Konflikt zwischen dem Anspruch der Richtigkeit der Entscheidung (Gefährdungsszenario 1) und dem Ziel der Nichtdiskriminierung durch ADM (Gefährdungsszenario 2).

Quantitative Begriffe der Gleichbehandlung bieten die Möglichkeit, das Vorhersageverhalten eines trainierten Modells genau zu untersuchen. Obgleich sie als direktes Regulierungsinstrument gegenwärtig nur bedingt geeignet sind, bilden sie potenziell nützliche Werkzeuge für die Gestaltung von Auditprozessen und bieten vielerlei Perspektiven für weitere interdisziplinäre Forschung in diesem Bereich. Die automatische Berücksichtigung solcher Metriken beim Erstellen von Modellen ist Gegenstand aktueller Forschung und eignet sich unserer Einschätzung nach noch nicht dazu regulativ aufgegriffen zu werden.

Es existiert eine Vielzahl von ML-Modellen verschiedener Komplexität mit verschiedenen Lernkapazitäten und Graden der Erklärbarkeit. Wenn nachvollziehbare Verfahren verwendet werden (wie z.B. die in der Praxis verbreitete logistische Regression), ist eine direkte Kontrolle des Modells prinzipiell möglich. Je nach Problem und Daten können diese jedoch komplexeren, weniger transparenten Modellen in Genauigkeit und Kapazität unterlegen sein. Komplexere Modelle erfordern entsprechend geeignete Prüfungskriterien. Es gilt daher die verschiedenen technischen Möglichkeiten der Modellinterpretation und -prüfung durch Testdaten im Rahmen von Test- und Auditprozessen sachgerecht mit den rechtlichen Anforderungen zu verknüpfen (siehe dazu die Diskussion in Kapitel 5.4.3).

Die Analyse eines konkreten ADM-Systems muss notwendigerweise in mehrere Ebenen unterteilt werden:

1. Analyse des Gesamtprozesses zur (Weiter-)Entwicklung eines ADM-Systems
2. Analyse des trainierten Modells
3. Analyse der Entscheidung auf Instanzebene

Diese Differenzierung ermöglicht einen holistischen Blick auf die komplexen Vorgänge und Abläufe, die bei ADM ineinandergreifen. Auf Basis dieses Verständnisses können auch unterschiedliche Maßnahmen identifiziert werden, die die Transparenz von ADM erhöhen und die Grundlage für den Nachweis von Fairness und Diskriminierung darstellen. Es zeigt sich, dass es nicht ausreicht, auf eine Ebene alleine abzustellen, um Entscheidungen nachvollziehen zu können bzw. um mögliche Ursachen für zu hinterfragende Entscheidungen zu erkennen und nachzuweisen. So können z.B. Biases in erhobenen Daten, wie oben bereits erwähnt wurde, zu Entscheidungen auf der Instanzebene führen, die diskriminierend sind. Von einem mathematisch-rationalen Standpunkt aus betrachtet ist dieses Verhalten eines maschinell trainierten Verfahrens durchaus nachvollziehbar und der Effekt, der sich auf der Instanzebene beobachten lässt, hat seine Ursache in einem der ersten Schritte innerhalb des Prozesses. Vermeidungsstrategien müssen daher dort bereits ansetzen, um effektiv und zielführend zu sein. In einer Arbeit von 2016 bringen Goodman und Flexman diesen Umstand auf den Punkt: *„machine learning depends upon data that has been collected from society, and to the extent that society contains inequality, exclusion or other traces of discrimination, so too will the data.“*[99]

Als Ergänzung zur Analyse des Gesamtprozesses und der darin vollzogenen Teilschritte haben wir im vorliegenden Kapitel zwei weitere Methoden diskutiert: Auditierung und Testen. Das Auditieren ist grundsätzlich geeignet, um die Transparenz der Entscheidungsstrukturen, die einem ADM-System zugrunde liegen, zu erhöhen. Es existieren jedoch unterschiedliche Auditierungsmethoden, die jeweils Vor- und Nachteile haben. Das klassische Code-Audit erfordert die Offenlegung des gesamten Softwareprogramms, sodass es für IT-Experten möglich ist, das System zur Gänze zu analysieren und zu erfassen. Je nach verwendetem maschinellem Lernverfahren können relevante Attribute sofort erkannt und Aussagen über eine mögliche problematische Diskriminierung getroffen werden. Dies gilt vor allem für Lernverfahren, die auf deduktiven und regelbasierten Methoden beruhen, z.B. Entscheidungsbäume. Für komplexere Lernverfahren, wie etwa neuronale Netze, kann die Code-Einsicht dennoch sehr erhellend sein, obwohl sich das trainierte Modell einer Interpretation durch den Menschen entzieht. An dieser Grenze können dann Testverfahren aufschlussreich sein. Diese können sehr effizient und schnell durchgeführt werden, sofern die Testfälle definiert sind. Allerdings ist die Definition von Testfällen stark von der jeweiligen Anwendung abhängig und in manchen Fällen nicht ohne weiteres möglich. So kann beispielsweise die Bonität von fiktiven Personen in Testdaten nicht geschätzt werden. Als mögliche Lösung kann der Einsatz von *Metamorphic Testing* dienen. Dabei wird nicht mehr das konkrete Ergebnis einer Klassifizierung betrachtet, sondern die Veränderung und das Verhältnis zwischen zwei Bewertungen. Es wird also geprüft, ob ein sensibles Attribut, z.B. Geschlecht oder Alter, einen signifikanten Unterschied in der Bewertung der Bonität macht oder nicht. Wir haben dieses Vorgehen anhand konkreter Fallbeispiele illustriert.

[99] Goodman/Flaxman, 2017.

Die hier angestellten Überlegungen bieten die technische Grundlage und einen Überblick über die Komplexität der Herausforderung, die sich bei der Interpretation und Erklärbarkeit von ADM ergibt. Sie zeigen Möglichkeiten auf, um sich der Thematik strukturiert zu nähern und für die Diskriminierung relevante Fragestellungen zu differenzieren. Sie spielen daher auch als Grundlage für die im nächsten Kapitel folgende juristische Betrachtung eine wichtige Rolle

5 Algorithmische Entscheidungen aus rechtlicher Sicht

5.1 Rechtsfragen algorithmischer Beurteilung von Personen

5.1.1 Stand der Diskussion und Problemlagen

5.1.1.1 Algorithmische Beurteilung von Personen in der juristischen Diskussion

Die Diskussion der spezifischen Rechtsfragen algorithmischer Entscheidungen hat erst jüngst begonnen. Umfassende Untersuchungen und Kategorisierungen der Problemlagen algorithmischer Entscheidungen aus rechtlicher Sicht liegen, soweit erkennbar, für das deutsche Recht nicht vor, was angesichts der Vielfalt der Fragestellungen und der Dynamik der Entwicklung verständlich ist. In der bereits genannten Untersuchung von Martini werden als Risiken die „Monopolisierung von Markt- und Meinungsmacht", „Intransparenz" und „Diskriminierung" genannt, aber ersichtlich nicht als systematische oder gar abschließende Problemanalyse verstanden.[100] Auch in anderen Rechtsordnungen werden eher einzelne Probleme beschrieben (dazu unten Kapitel 6). Vielfach werden in der juristischen Diskussion dieselben oder ähnliche Themen genannt. So wird auf die Rolle von Facebook im US-Wahlkampf verwiesen, die Bedeutung des Wohnorts für Kreditentscheidungen oder die Relevanz des Vornamens in automatisierten Bewerbungsverfahren. Häufig wird auch auf die in den USA eingesetzte Software zur Ermittlung der Rückfallwahrscheinlichkeit von Straftätern abgestellt. Ein umfassender Befund oder gar eine einheitliche Problembeschreibung besteht auch in der internationalen Perspektive nicht.

Eine intensive rechtliche Auseinandersetzung mit den Herausforderungen der Beurteilung von Personen durch Maschinen existiert bisher nicht. Auch fehlt es an einer systematischen Beschreibung deren spezifischen rechtlichen Probleme.

Als eine spezifische Problemlage des Einsatzes algorithmischer Entscheidungen kann wohl die Steigerung des Ausmaßes der Beurteilung von Menschen identifiziert werden. Dies gilt sowohl hinsichtlich der Intensität als auch des Umfang von Beurteilungen.

5.1.1.2 Beurteilung von Menschen in neuen Bereichen

So lässt sich beobachten, dass mit der Nutzung algorithmischer Entscheidung eine Beurteilung in Bereichen eintritt, in denen sie bisher nicht vorgenommen wurde. So werden Güter des Alltags im Massengeschäft traditionell ohne Beurteilung des Käufers verkauft. Im Supermarkt steht die Ware für jeden Käufer mit gleichen Merkmalen gleichermaßen zur Verfügung.

Dies ändert sich etwa durch *dynamic pricing*, soweit der Preis einer Ware abhängig von der Einschätzung einer Person, etwa deren Zahlungsbereitschaft, festgelegt wird. Individuelle Preise als solches sind nichts Neues und das Wesensmerkmal aller Märkte, in denen der Preis durch individuelle Verhandlung festgelegt wird. Es ist aber ein neues Phänomen, wenn

[100] Vgl. Martini 2017.

die individuelle Preisfindung in den Prozess von Massengeschäften ohne expliziten Verhandlungsprozess übertragen wird.[101] Grund dieser Entwicklung ist offenbar die Absenkung der Transaktionskosten durch den Einsatz von Maschinen auf Seiten des Anbieters der Ware, die die individuelle Bepreisung für den Verkäufer wirtschaftlich lohnend macht.

5.1.1.3 Gesteigerte Intensität und einseitige Steuerung der Beurteilung

Die Möglichkeiten algorithmischer Entscheidungen können auch zu einer gesteigerten Intensität der Beurteilung führen. So können in eine algorithmische Beurteilung von Personen unter Umständen sehr viel mehr Informationen einfließen, als sie etwa einem menschlichen Entscheider zur Verfügung stehen. So hat ein menschlicher Verkäufer etwa im Massengeschäft meist nur wenig Informationen über einen Kaufinteressenten. Dies kann bei maschinellen Entscheidungen, wenn auf umfassende Datenbestände zugegriffen werden kann, entscheidend anders sein. Auch bei einer Personalauswahl können bei algorithmischen Entscheidungen unter Umständen weit mehr Informationen herangezogen werden, als es bei menschlichen Sachbearbeitern der Fall wäre.

Ein Problem der Beurteilung durch Algorithmen kann sich auch durch die Verengung der Beurteilungsgrundlage, verbunden mit der einseitigen Steuerung der Beurteilung, ergeben. ADM-Systeme machen Vorhersagen auf der Basis der zur Verfügung stehenden Daten und nur dieser Daten. Während der Verbraucher bei unmittelbarer Kommunikation mit einem menschlichen Beurteiler typischerweise initiativ werden kann, wird ihm diese Dimension der Verhandelbarkeit beim Einsatz von ADM-Systemen entzogen. Vielmehr wird es zur Aufgabe des Kunden, sich ADM-konform zu verhalten. Ein eindrucksvolles Beispiel hierfür ist die vor allem in den USA bekannte Kreditscore-Optimierung, bei der spezialisierte Berater durch gezielte Maßnahmen, z.B. die Umschichtung von Schulden, die errechnete Kreditwürdigkeit erhöhen.

5.1.1.4 Zwischenergebnis

Die hier vermuteten spezifischen Problemlagen, die Ausweitung der Beurteilung von Personen sowohl im Anwendungsbereich als auch in der Intensität durch ADM-Systeme, werden insbesondere rechtlich relevant, wenn die Beurteilung fehlerhaft ist oder wenn durch die Beurteilung Ungleichgewichte entstehen.

Diese Studie nimmt daher vor allem die Probleme fehlerhafter Beurteilungen von Personen in den Blick. Außerdem werden etwaige Ungleichgewichtslagen durch Einsatz von ADM, insbesondere im Verhältnis zwischen Unternehmern und Verbrauchern, untersucht.

5.1.2 Ungleichgewicht zwischen Verbraucher und Unternehmer durch algorithmische Entscheidungen

5.1.2.1 Verdacht auf problematische Ungleichgewichtslagen

Algorithmische Entscheidungen werden häufig im Verhältnis zwischen Unternehmern und Verbrauchern, und regelmäßig ausschließlich durch den Unternehmer, eingesetzt. Dies löst den Verdacht auf Entstehung oder Vertiefung einer Ungleichgewichtslage zwischen

[101] Auf Probleme durch *dynamic pricing* weist etwa die BaFin in ihrer Studie „Big Data trifft auf künstliche Intelligenz“, Ziff. 6.2.3.1 (S. 182) hin.

Unternehmer und Verbraucher aus.[102] Insbesondere liegt es nahe, dass eine Informationsasymmetrie zu Lasten des Verbrauchers entsteht oder erweitert wird.

Ein anderer Aspekt betrifft die Ausnutzung eines individuellen Bedürfnisses des Verbrauchers an einer Leistung durch den Unternehmer. Dieser Aspekt ist eng mit dem Problem der Informationsasymmetrie verknüpft, aber nicht damit identisch. So kann beispielsweise der Unternehmer dem Verbraucher aufgrund Kenntnis dessen zwingenden Bedürfnisses an der nachgefragten Leistung einen individuellen – höheren – Preis festlegen, ohne dass der Verbraucher – selbst bei vollständiger Aufklärung – diesem Verhalten entgehen kann.

Ein einfaches Beispiel: Wenn der Betreiber eines Fußballstadions weiß, dass ein Fußballfan ein bestimmtes Spiel unbedingt im Stadion sehen möchte, und deswegen den Ticketpreis für diesen Fan entsprechend anhebt, nützt dem Fußballfan etwaiges Wissen darüber, dass ihm aufgrund seiner konkreten Bedürfnislage ein erhöhter Preis angeboten wird, nichts. Dasselbe gilt bei der Buchung einer Flugreise.

In derartigen Situationen sind Fragen der Gleichbehandlung von Bedeutung (siehe dazu Kapitel 5.1.2.3).

5.1.2.2 Informationsasymmetrie

Algorithmische Beurteilungen von Menschen sind eng mit Problem der Informationsasymmetrie verknüpft. Informationsasymmetrien im weiteren Sinne als unterschiedliches Wissen zwischen Partnern in rechtsgeschäftlichen oder sozialen Beziehungen können problematisch sein, sind zugleich aber notwendige Grundlage einer arbeitsteiligen Gesellschaft. Entsprechend differenziert werden Informationsasymmetrien rechtlich adressiert, insbesondere durch punktuelle Informations- und Aufklärungspflichten oder gar weitergehende Treuepflichten in Vertragsverhältnissen (Arzt, Rechtsanwalt).

Beim Einsatz von ADM-Systemen können Informationsasymmetrien im Verhältnis von Unternehmern und Verbrauchern aus rechtlicher Sicht von unterschiedlicher Bedeutung sein. So kann der Unternehmer aufgrund der algorithmischen Beurteilung mehr Informationen über den Verbraucher und dessen Interesse an der Leistung haben als umgekehrt der Verbraucher über die Leistung des Unternehmers und Alternativen. Damit eng verbunden ist der Aspekt, dass es dem Verbraucher an Kenntnis darüber fehlen kann, was der Unternehmer über ihn und seine Verhältnisse weiß.

Diese Aspekte sind aus rechtlicher Sicht unterschiedlich zu bewerten, etwaige rechtliche Gegenmaßnahmen können durchaus andersartig sein: So ist etwa die Unkenntnis des Verbrauchers über die Merkmale der angebotenen Leistung und Alternativen durch entsprechende Information über die Leistung des Unternehmers und die Möglichkeit zum Vergleich zu beheben. Die Erteilung dieser Information kann rechtlich etwa durch produktbezogene Informations- und Aufklärungspflichten gesichert werden, wie sie im klassischen Verbraucherschutzrecht üblich sind. Dagegen entspricht das Interesse des Verbrauchers an der Kenntnis der über ihn verwendeten Informationen klassischen datenschutzrechtlichen Anliegen, die im Hinblick auf Verbraucherschutz gegenüber ADM zu analysieren und gegebenenfalls fortzuentwickeln sind.

[102] Martini (2017) verweist insoweit auf eine „Markt-Macht-Spirale“, ohne diese allerdings näher zu spezifizieren.

Informationsasymmetrien im Verhältnis zwischen Unternehmen und Verbrauchern sind Gegenstand intensiver Forschung verschiedener Disziplinen und Gegenstand von Gesetzgebung. Ob und inwieweit algorithmische Entscheidungen typische Problemlagen der Informationsasymmetrie betreffen, bedarf weiterer Forschung, insbesondere der Vernetzung mit der allgemeinen Forschung zu Informationsasymmetrie im Verhältnis zwischen Unternehmern und Verbrauchern, die im Rahmen dieser Studie nicht erfolgen kann. Die Studie beschränkt sich daher auf Aspekte, die unmittelbar mit ADM verbunden sind, etwa die Reichweite von datenschutzrechtlichen Informationspflichten und Auskunftsansprüchen im Kontext von ADM.

5.1.2.3 Gleichbehandlung und algorithmische Entscheidungen

5.1.2.3.1 Gleichbehandlung als Fragestellung algorithmischer Entscheidungen

Algorithmische Entscheidungen stellen das Recht nicht zuletzt unter dem Aspekt der Gleichbehandlung vor Herausforderungen. So wirft etwa das *dynamic pricing* die Frage auf, ob und in welchem Umfang die Gleichbehandlung in Bezug auf Preise rechtlich geboten ist oder sein soll. Der Vertragsfreiheit sind bei der Preisbildung insoweit nur eingeschränkt Grenzen gesetzt, zivilrechtlich etwa durch § 138 BGB, vor allem aber durch das Kartellrecht. Ob der Einsatz von *dynamic pricing* eine Gefahr für den Wettbewerb ist, die durch das Kartellrecht erfasst werden kann, wird derzeit zwar durchaus untersucht. Es ist aber noch völlig offen, wie weit die Einschränkungen aus kartellrechtlicher Sicht reichen.

Ähnliche Fragen stellen sich bei der automatisierten Personalauswahl. Zwar greifen insoweit Diskriminierungsverbote ganz unterschiedlicher Art und Rechtsgrundlagen. Davon abgesehen sieht das Recht keine umfassende Einschränkung ungleicher oder gar „ungerechter“ Personalauswahl vor.

Es ist zu vermuten, dass der Gegensatz zwischen dem in der Vertragsfreiheit zum Ausdruck kommenden grundrechtlich geschützten Gestaltungsfreiheit und dem Schutz Betroffener in Ungleichgewichtslagen nicht durch einen allgemeinen, übergreifenden Grundsatz zu lösen sein wird, sondern den Spezifika der jeweiligen Interessenlagen folgend in den jeweiligen Einsatzbereichen algorithmischer Entscheidungen spezifisch auszugleichen ist.

5.1.2.3.2 Gleichbehandlungsgrundsatz und individuelle Beurteilung

Eine allgemeine Frage, die sich bei zahlreichen Formen algorithmischer Entscheidungen stellt, ist, inwieweit die Verwendung von Information sowie die algorithmische Beurteilung in Alltagssituationen gerechtfertigt sind und in welchem Maße ein Anspruch auf Gleichbehandlung existiert.

Auf verfassungsrechtlicher Ebene besteht ein Grundrecht auf Gleichbehandlung nach Art. 3 GG des Einzelnen gegenüber dem Staat. Im Verhältnis zwischen Privaten wirkt der Gleichbehandlungsgrundsatz indes nicht unmittelbar, kann aber mittelbar einwirken. Auf der Ebene des Zivilrechts findet der Gleichbehandlungsgrundsatz, soweit nicht spezifisch geregelt, durch allgemeine Grundsätze des Vertragsrechts, insbesondere des § 242 BGB (Treu und Glauben), Berücksichtigung. Aus diesen Grundsätzen folgt etwa, dass ein willkürlicher Ausschluss eines Vertragspartners von einer Leistung unzulässig sein kann.

Die Reichweite des Gleichbehandlungsgebots in modernen Formen der Leistungsdistribution ist freilich noch nicht geklärt. Insoweit besteht erheblicher Forschungsbedarf.

5.1.2.3.3 Recht auf Gleichbehandlung bei struktureller Überlegenheit?

Für die Frage nach der Bedeutung des Gleichbehandlungsgrundsatzes für algorithmische Entscheidungen lassen sich möglicherweise Erkenntnisse aus der Bedeutung des Gleichbehandlungsgrundsatzes in Massengeschäften ziehen.

Insoweit ist ein aktueller Beschluss des Bundesverfassungsgerichts von Interesse, der die Bedeutung des Gleichbehandlungsgrundsatzes für die Verhängung eines Stadionverbots zum Gegenstand hatte.[103] Ausgangspunkt war die Verfassungsbeschwerde eines Fußballfans gegen ein zivilrechtliches Urteil, das das gegen ihn verhängte Stadionverbot bestätigte.[104]

Das BVerfG weist in dem Beschluss auf die allgemeinen Grundsätze des Gleichbehandlungsgrundsatzes hin, insbesondere auf den Umstand, dass Art. 3 GG kein Gleichbehandlungsgebot für die Rechtsbeziehungen zwischen Privaten enthält und sich ein solches auch nicht aus den Grundsätzen der mittelbaren Drittwirkung von Grundrechten ergibt. Es fasst dies wie folgt zusammen: „Ein allgemeiner Grundsatz, wonach private Vertragsbeziehungen jeweils den Rechtfertigungsanforderungen des Gleichbehandlungsgebots unterlägen, folgt demgegenüber aus Art. 3 Abs. 1 GG auch im Wege der mittelbaren Drittwirkung nicht.“[105]

Grundrechtliche Anforderungen bestehen aber, wie das BVerfG im Anschluss daran betont, in bestimmten Situationen.[106] Sodann entwickelt das BVerfG den folgenden Grundsatz:

„Maßgeblich für die mittelbare Drittwirkung des Gleichbehandlungsgebots ist dessen Charakter als einseitiger, auf das Hausrecht gestützter Ausschluss von ***Veranstaltungen, die aufgrund eigener Entscheidung der Veranstalter einem großen Publikum ohne Ansehen der Person geöffnet werden*** *und der für die Betroffenen in erheblichem Umfang über die Teilnahme am gesellschaftlichen Leben entscheidet. Indem ein Privater eine solche Veranstaltung ins Werk setzt, erwächst ihm von Verfassungswegen auch eine besondere rechtliche Verantwortung. Er darf seine hier aus dem Hausrecht – so wie in anderen Fällen möglicherweise aus einem Monopol oder* ***aus struktureller Überlegenheit – resultierende Entscheidungsmacht*** *nicht dazu nutzen, bestimmte Personen ohne sachlichen Grund von einem solchen Ereignis auszuschließen.*“[107]

In der Folge misst das BVerfG die angegriffene Entscheidung des BGH an diesem Grundsatz und weist die Verfassungsbeschwerde schließlich ab.[108]

Diese Überlegungen des BVerfG sind für die rechtliche Erfassung der strukturellen Überlegenheit in algorithmischen Entscheidungen von großem Interesse. Zunächst bestätigt das BVerfG, dass ein Recht auf Gleichbehandlung zwischen Privaten nur in bestimmten Konstellationen besteht. Grundlage für ein Eingreifen des Gleichbehandlungsgrundsatzes kann die Eröffnung einer Veranstaltung für ein großes Publikum ohne Ansehung der Person sein. Aus dem Gleichbehandlungsgrundsatz folgt in derartigen Fällen, dass eine aus struktureller Überlegenheit resultierende Entscheidungsmacht einer rechtlichen Kontrolle unterliegt.

[103] BVerfG, Beschl. v. 11.4.2018, 1 BvR 3080/09.
[104] BVerfG, Beschl. v. 11.4.2018, 1 BvR 3080/09, Rn. 9 ff.
[105] BVerfG, Beschl. v. 11.4.2018, 1 BvR 3080/09, Rn. 40.
[106] BVerfG, Beschl. v. 11.4.2018, 1 BvR 3080/09, Rn. 41.
[107] BVerfG, Beschl. v. 11.4.2018, 1 BvR 3080/09, Rn. 41 (Hervorhebungen vom Verfasser).
[108] BVerfG, Beschl. v. 11.4.2018, 1 BvR 3080/09, Rn. 49 ff.

5.1.2.3.4 Algorithmische Entscheidungen und strukturelle Überlegenheit

Inwieweit die aktuelle Entscheidung des BVerfG und die dort genannten Grundsätze eine Bedeutung für algorithmische Entscheidungen haben, ist offen. Das BVerfG hat seine Entscheidung ganz ausdrücklich auf die spezifische Situation des Stadionverbots, die seinem Beschluss zugrunde lag, beschränkt.

Es spricht aber einiges dafür, dass die Entscheidung Ausdruck eines allgemeinen Grundsatzes ist, da das BVerfG das im konkreten Fall zu beurteilende Hausrecht in eine Reihe mit einem Monopol und einer strukturellen Überlegenheit stellt. Das BVerfG verweist in seinem Beschluss auf zwei Voraussetzungen eines Gleichbehandlungsgebots: die Existenz einer einem großen Publikum geöffneten Veranstaltung und die Bedeutung der Teilnahme an der Veranstaltung für die Teilnahme am gesellschaftlichen Leben.

Eine zur Gleichbehandlung verpflichtende Überlegenheit wird traditionell bei Monopolen und Angewiesenheit auf die jeweilige Leistung angenommen. Im Hinblick auf den schwierigen Begriff des Monopols kann ein Gleichbehandlungsgebot auch bei „situativen Monopolen“ entstehen, sofern das Gegenüber auf die jeweilige Leistung angewiesen ist. Voraussetzung für einen Gleichbehandlungsanspruch dürfte aber, wie das BVerfG in seinem aktuellen Beschluss betont, die Eröffnung eines Verkehrs für ein allgemeines Publikum sein.

Das BVerfG nennt in der zitierten Passage das Hausrecht neben einem Monopol und der strukturellen Überlegenheit als Grundlage für eine einseitige Gestaltungsmacht und deutet damit an, dass die strukturelle Überlegenheit als solche eine Grundlage für ein Gleichbehandlungsgebot darstellen kann. Der Begriff der strukturellen Überlegenheit wird auch zur Bezeichnung des Verhältnisses von Unternehmern und Verbrauchern verwendet. Es ist aber offensichtlich, dass das BVerfG an ein Gleichbehandlungsgebot wesentlich höhere Voraussetzungen setzt als die bloß abstrakte, rollenspezifische Überlegenheit des Unternehmers im Verhältnis zum Verbraucher. Man wird vielmehr anzunehmen haben, dass eine strukturelle Überlegenheit ein Gleichbehandlungsgebot nur auslösen kann, wenn die für das Gleichbehandlungsgebot typischen Merkmale – die Öffnung der Leistung für einen breiten Verkehr, die Angewiesenheit auf die Leistung sowie eine einseitige Verfügungsmacht des Anbieters über die Leistung – in vergleichbarer Weise wie bei einem Monopol vorliegen.

Unter diesen Voraussetzungen wird man wohl einen allgemeinen Grundsatz dahin formulieren können, dass die aus struktureller Überlegenheit resultierende, einer Monopolstellung betreffend einer wesentlichen Leistung gleichkommende, Entscheidungsmacht einen Vertragspartner verpflichtet, potenzielle Vertragspartner nicht ohne sachlichen Grund ungleich zu behandeln.

Nimmt man einen solchen Grundsatz an, ist offensichtlich, dass der Einsatz von ADM-Systemen nicht per se zu einer derartigen zur Gleichbehandlung verpflichtenden strukturellen Überlegenheit führt.

Andererseits liegt es jedoch nahe, dass der Einsatz von ADM-Systemen in Massengeschäften unter dem Gesichtspunkt einer strukturellen Überlegenheit zu einem Gleichbehandlungsgebot führen kann. Dies wird aber nur dann in Betracht kommen, wenn die algorithmische Beurteilung von Personen genutzt wird, um die Angewiesenheit einer Person auf die Leistung und damit das Kernelement der strukturellen Überlegenheit im Sinne des Gleichbehandlungsgebots zu erzeugen, zu verstärken oder auszunutzen.

In vielen Einsatzbereichen algorithmischer Entscheidungen wird diese Situation nicht gegeben sein. Wenn etwa ADM-Systeme genutzt werden, um die Kosten für die

Bearbeitung von Massengeschäften zu senken, liegt darin noch kein Einsatz zum Zweck der Erzeugung oder Ausnutzung einer Angewiesenheit auf eine Leistung. Wird das Verfahren aber genutzt, um Vertragspartner zu identifizieren, die in besonderer Weise auf die Leistung angewiesen sind, um von diesen entsprechend höhere Preise zu verlangen, wird strukturelle Überlegenheit genutzt.

Damit wird offenbar, dass sich in Fällen wie beispielsweise *dynamic pricing* möglicherweise Einschränkungen aus dem verfassungsrechtlichen Gleichbehandlungsgebot herleiten lassen. Indes sind die Voraussetzungen und Rechtsfolgen derzeit noch völlig unklar.

5.1.2.4 Zwischenergebnis

Algorithmische Entscheidungen werfen, über die Konstellation der Diskriminierung oder sonst rechtlich fehlerhafter Entscheidungen hinaus, Fragen in vielfachen Problemfeldern auf.

5.1.2.4.1 Strukturelle Überlegenheit des Unternehmers durch ADM

Als ein Zwischenergebnis ist festzustellen, dass durch maschinelle Beurteilung vermutlich ein gesteigertes Ungleichgewicht zwischen Unternehmern und Verbrauchern entsteht. Im Vergleich zu traditionellen Situationen der Leistungserbringung, etwa dem Vertrieb von Produkten in einem Verkaufslokal ohne Einsatz von KI, stehen dem Unternehmer mehr Informationen über Verbraucher zur Verfügung als früher, ohne dass der Verbraucher diese kennt oder dies beeinflussen kann. Der Unternehmer kann diese Informationsüberlegenheit in vielfältiger Weise, etwa zum *dynamic pricing*, nutzen. Ob dem Verbraucher insoweit hinreichende Möglichkeiten zur Verfügung stehen, seine Interessen zu wahren, ist jedenfalls zweifelhaft.

Als These ergibt sich damit, dass der Einsatz von ADM zu einem strukturellen Ungleichgewicht zwischen Unternehmern und Verbrauchern führen kann. Der genaue Umfang und die Auswirkungen des hierdurch entstehenden Ungleichgewichts sind noch recht wenig erforscht und können im Rahmen dieser Studie nicht geklärt werden.

5.1.2.4.2 Gleichbehandlungsgebote wegen struktureller Überlegenheit?

Als ein weiteres Zwischenergebnis ist festzustellen, dass der Einsatz von ADM-Systemen in bestimmten Situationen wohl zu einem Gleichbehandlungsgebot unter dem Gesichtspunkt einer strukturellen Überlegenheit führen kann. Ein solches Gleichbehandlungsgebot, das aus dem verfassungsrechtlichen Gleichheitsgrundsatz (Art. 3 GG) abgeleitet werden kann, wird in Betracht kommen, soweit die algorithmische Beurteilung von Personen genutzt wird, um die Angewiesenheit auf die Leistung und damit das Kernelement der strukturellen Überlegenheit im Sinne des Gleichbehandlungsgebots zu erzeugen, zu verstärken oder auszunutzen.

Die Voraussetzungen eines solchen aus der Verfassung ableitbaren Gleichbehandlungsgebots in algorithmischen Entscheidungen sind derzeit jedoch noch sehr unklar. Die damit angesprochenen Fragen können im Rahmen dieser Studie nicht geklärt werden und bedürfen weiterer Forschung (dazu unten Kapitel 8).

5.2 Fehler algorithmischer Beurteilung von Menschen

Gegenstand der Studie ist insbesondere die Beurteilung von Personen durch Maschinen, aufgrund derer eine Entscheidung getroffen wird. So beruht etwa die Entscheidung über die Bereitschaft zum Abschluss eines Kreditvertrags oder die Auswahl bestimmter Konditionen für einen Kreditvertrag auf einer Beurteilung der Kreditwürdigkeit des Kreditnehmers.

Eine „Beurteilung" in diesem Sinne ist regelmäßig Bestandteil einer Auswahlentscheidung. Wenn beispielsweise ein Vermieter zwischen zwei Mietinteressenten auswählt, liegt der Auswahlentscheidung typischerweise eine Beurteilung der Mietinteressenten in Bezug auf ihre Eignung als Mieter zugrunde. Eine solche Beurteilung liegt nach diesem Verständnis nur dann nicht vor, wenn die Auswahlentscheidung vollständig vom Zufall abhängig gemacht wird.

Staatliche Aufgaben in Bezug auf den Einsatz von ADM-Systemen können nicht zuletzt hinsichtlich der Sicherung der Qualität der Entscheidung bestehen. Wie etwa Wischmeyer in seiner aktuellen Untersuchung aufzeigt, kann sich staatlicher Handlungsbedarf zur Qualitätssicherung insbesondere zum Schutz von Grundrechten, darüber hinaus aber auch zum Schutz anderer Verfassungswerte, etwa Demokratie und Rechtsstaatlichkeit, ergeben.[109] Die Notwendigkeit zur rechtlichen Regelung einer maschinellen Beurteilung kann sich vor allem aus problematischen Beurteilungen ergeben, die nachfolgend als „fehlerhafte" Beurteilungen bezeichnet werden. Damit ist von entscheidender Bedeutung, unter welchen Voraussetzungen die automatisierte Beurteilung eines Menschen als problematisch oder fehlerhaft anzusehen ist.

Eine allgemeingültige Definition des Begriffs der Fehlerhaftigkeit bei der Beurteilung von Personen existiert, soweit ersichtlich, nicht. Dies dürfte wesentlich auf den Umstand zurückzuführen sein, dass die Bewertung einer Beurteilung als „fehlerhaft" in mehrfacher Hinsicht von der Perspektive und den Zielen des Bewertenden abhängt.

So kann beispielsweise die Bevorzugung von Bewerbern einer bestimmten Altersgruppe im Rahmen der Personalauswahl aus Sicht des Auswählenden richtig sein, etwa weil er die altersmäßige Homogenität des Personals für wichtig hält. Aus Sicht des Bewerbers hingegen kann es sich um Altersdiskriminierung handeln; aus Sicht eines Vorgesetzten, der altersgemischte Teams bevorzugt, wiederum um eine fachlich unqualifizierte Entscheidung. Gerade im Hinblick auf den Gegenstand der Studie, die algorithmische Beurteilung, scheint es wichtig, diese Relativität des Fehlerbegriffs in den Fokus zu nehmen, da die Funktion der Maschine häufig indifferent ist. Nach dem Zweck dieser Studie kann man vereinfachend zwischen zwei Perspektiven unterscheiden: Fehler als Abweichung von eigenen Zielen des Entscheiders und Fehler als Abweichung von normativen Anforderungen.

5.2.1 Abweichung von eigenen Zielen des Entscheiders

Eine Beurteilung kann zum einen im Hinblick auf die vom Beurteilenden selbst gewählten Ziele „fehlerhaft" oder „mangelhaft" sein. Ein solcher Mangel der Beurteilung liegt dann vor, wenn sie in der Entscheidungsfindung oder im Entscheidungsergebnis von den vom Beurteilenden selbst als maßgeblich angenommenen Maßstäben abweicht. Dies wäre beispielsweise der Fall, wenn der Beurteilende eine von ihm selbst als relevant angesehene Information übersieht oder sie ihm nicht zur Verfügung stand. Im Gefährdungsszenario

[109] Wischmeyer, AöR 2018, 1, 23 f.

„Unrichtiges Kreditscoring“ (vgl. Kapitel 2.2) liegt häufig, wenngleich nicht notwendig,[110] ein solcher Beurteilungsmangel vor.

5.2.2 Fehler als Abweichung von normativen Anforderungen

Eine Beurteilung kann zum anderen im Hinblick auf die von Dritten vorgegebenen Ziele fehlerhaft sein. Diese Fallgruppe soll hier hinsichtlich normativer Vorgaben adressiert werden.

Ein Fehler der Beurteilung liegt nach diesem Maßstab vor, wenn sie in der Entscheidungsfindung oder im Entscheidungsergebnis von normativ vorgegebenen Maßstäben, insbesondere von Anforderungen rechtlicher oder sonstiger (z.B. ethischer) Normen, abweicht. Dies wäre etwa der Fall, wenn ein rechtlich unzulässiges Entscheidungskriterium verwendet wird, wie es bei der Diskriminierung der Fall ist, oder wenn der Beurteilende eine Information übersieht, die nach normativen Anforderungen zu berücksichtigen wäre.

Für die Zwecke der Studie wird unter dem Begriff des „Fehlers“ einer Beurteilung die Abweichung von rechtlichen Anforderungen an die Entscheidungsfindung oder das Entscheidungsergebnis verstanden.

Ein Fehler in diesem Sinne kann in allen drei Gefährdungsszenarien existieren. Soweit eine rechtliche Verpflichtung zu einem zutreffenden Kreditscoring besteht, liegt bei einem unzutreffenden Scoring ein Fehler der Beurteilung vor. Eine fehlerhafte Beurteilung liegt im Fall der Diskriminierung vor, soweit man diese als rechtlich unzulässige Ungleichbehandlung versteht. Im Gefährdungsszenario „Intransparenz“ gilt dies ebenfalls, wenn und soweit eine rechtliche Transparenzpflicht verletzt wird.

Eine weitere für die rechtliche Regelung der Beurteilung wichtige Unterscheidung ist im Hinblick auf die Bewertung der Beurteilung nach der Perspektive und dem Wissensstand der Person zu treffen, die die Beurteilung als fehlerhaft oder fehlerfrei zu bewerten hat.

Diese Unterscheidung wird in der rechtlichen Diskussion meist mit dem Begriffspaar „ex ante“ und „ex post“ gekennzeichnet. Dabei wird mit dem Begriff „ex ante“ meistens die Situation des Entscheidenden (hier: des Beurteilenden), insbesondere dessen Wissensstand und Erkenntnismöglichkeiten bezeichnet, wogegen mit dem Begriff „ex post“ die Perspektive eines mit Zusatzwissen ausgestatteten Dritten benannt wird, der die Beurteilung bewertet. Musterfall ist die Perspektive eines Gerichts, das über die Fehlerhaftigkeit einer Beurteilung zu entscheiden hat.

5.3 Arten von Fehlern

Eine allgemeingültige Systematik der Fehlerhaftigkeit algorithmischer Entscheidungen oder der Fehlerhaftigkeit der Beurteilung von Personen existiert, soweit ersichtlich, nicht. In seiner aktuellen Untersuchung zu Rechtsfragen von Algorithmen nennt Martini Intransparenz und

[110] Wenn es das ausschließliche Ziel des Kreditscorings ist, die Kreditvergabe an einen kreditunwürdigen Darlehensinteressenten zu vermeiden, steht die fehlerhafte Beurteilung als „kreditunwürdig“ nicht notwendig im Gegensatz zu den Zielen des Entscheiders. In der Praxis wollen Banken allerdings kreditwürdigen Darlehensnehmern einen Kredit geben und sind daher an einer in beide Richtungen zutreffenden Bewertung interessiert.

Diskriminierung als – erkennbar nicht abschließende – Fallgruppen von Problemen algorithmischer Entscheidungen.[111]

Für die Zwecke der Studie sollen die nachfolgend genannten Fallgruppen unterschieden werden:

- Unzulässigkeit der (algorithmischen) Beurteilung
- Intransparenz der Beurteilung
- Fehler der Entscheidungsfindung/Beurteilungsverfahren
- Fehler der Entscheidungsgrundlage
- Fehler bei Würdigung der Entscheidungsgrundlagen

5.3.1 Unzulässigkeit der (algorithmischen) Beurteilung

Ein Fehler einer Beurteilung kann bereits in deren Vornahme, Äußerung oder Verwendung als Entscheidungsgrundlage vorliegen. Äußerungsverbote sind häufig und ergeben sich etwa aus Geheimhaltungs- und Verschwiegenheitsgeboten. Soweit beispielsweise die ärztliche Schweigepflicht greift, wäre zwar nicht die Erstellung, aber jedenfalls die Mitteilung der Beurteilung eines Patienten an Dritte unzulässig.

Die bloße Vornahme einer Beurteilung durch natürliche Personen ist, soweit erkennbar, nicht geregelt, offensichtlich wegen der traditionellen Unzugänglichkeit menschlicher Denkvorgänge.

Im Hinblick auf die algorithmische Beurteilung kann schon die Vornahme oder Verwendung als Entscheidungsgrundlage unzulässig sein. Dieser Aspekt ist angesprochen, soweit etwa angenommen wird, in Verwaltungsverfahren oder in Strafverfahren[112] sei eine vollautomatisierte Entscheidung unzulässig. Im Anwendungsbereich eines solchen Verbots wäre eine gleichwohl vorgenommene Beurteilung fehlerhaft im hier genannten Sinne.

Dieser Gedanke (Verbot der algorithmischen Beurteilung) liegt auch etwa Art. 22 DSGVO zugrunde, der dem Betroffenen einen Anspruch auf eine Beurteilung durch eine natürliche Person gewährt.[113]

5.3.2 Intransparenz der Beurteilung

Eine Beurteilungsentscheidung kann bereits wegen Intransparenz fehlerhaft sein. Dies ist etwa dann der Fall, wenn die Entscheidung öffentlich zu erfolgen hat. Als Fallgruppe der Intransparenz lässt sich auch das Fehlen einer gebotenen Begründung verstehen.

5.3.3 Fehler der Entscheidungsfindung/Beurteilungsverfahren

Eine Entscheidung kann fehlerhaft sein, weil sie Anforderungen an das Verfahren der Entscheidungsfindung verletzt. Fehler des Verfahrens können sich auf die Sammlung der tatsächlichen Entscheidungsgrundlagen beziehen, die hier als eigene Fallgruppe genannt ist. Es gibt aber auch zahlreiche andere Anforderungen an ein Entscheidungsverfahren,

[111] Martini 2017.
[112] Siehe dazu Wischmeyer, AöR 2018, 1, 64, Fn. 260 mit Nachweisen zur US-amerikanischen Debatte.
[113] Siehe dazu unten Kapitel 5.5.

etwa die Gewährung der Gelegenheit zu Stellungnahmen oder die Auswahl des Beurteilenden (z.B. Befangenheit).

5.3.4 Fehler der Entscheidungsgrundlage

Als Fehler der Entscheidungsgrundlage sollen hier Fehler bei der Sammlung der Entscheidungsgrundlagen, insbesondere der für die Beurteilung zugrundeliegenden Tatsachenbasis, verstanden werden. Fehlerhaft kann die Entscheidungsgrundlage etwa durch Heranziehung unzulässiger Tatsachen sein, wie es bei Diskriminierung (vgl. dazu unten Kapitel 5.4) der Fall ist. Diese Fallgruppe ist aber wesentlich weiter und umfasst auch die Heranziehung unzutreffender Tatsachen oder eine unvollständige Sammlung von Tatsachen.

Zu beachten ist, dass nach dem für die Studie zugrunde gelegten Fehlerverständnis ein Fehler nur vorliegt, soweit das Recht die Heranziehung einer vollständigen oder fehlerfreien Tatsachenbasis verlangt, wie es insbesondere bei behördlichen oder gerichtlichen Verfahren grundsätzlich (oft eingeschränkt wie z.B. durch den Beibringungsgrundsatz im Zivilprozess etc.) der Fall ist. Im privatrechtlichen Bereich ist dies nicht durchgehend der Fall.

Als Fehler der Entscheidungsgrundlage können auch Fehler bei der Ermittlung der maßgeblichen Rechtsgrundlagen (z.B. ausländisches Recht) verstanden werden. Da die Abgrenzung zu Fehlern der Entscheidungsfindung insofern aber schwierig ist, sollen derartige Fehler im Rahmen dieser Studie nicht explizit angesprochen werden.

5.3.5 Fehler bei Würdigung der Entscheidungsgrundlagen

Fehler können sich bei der Beurteilung oder Würdigung der Entscheidungsgrundlagen ergeben. So kann einem Umstand ein objektiv nicht angemessenes Gewicht beigemessen werden, etwa wenn bei der Personalauswahl ausschließlich die Examensnote herangezogen und die berufliche Erfahrung nicht bewertet wird oder wenn ein Rechenfehler bei der Addition von Punkten für die Gesamtbewertung unterläuft.

5.4 Diskriminierung

Zu den Fehlern, die bei einer (algorithmischen oder menschlichen) Entscheidung auftreten können, gehört die unzulässige Diskriminierung – also das zweite Gefährdungsszenario, das wir in der vorliegenden Studie betrachten. Der Begriff der Diskriminierung wird im allgemeinen Sprachgebrauch und im Recht in verschiedenen Kontexten verwendet. Darunter wird teils eine „benachteiligende oder bevorzugende Ungleichbehandlung“[114] verstanden; in dieser Bedeutung wird der Begriff auch etwa in den Wirtschaftswissenschaften verwendet, wo der Terminus der Preisdiskriminierung zunächst wertneutral ist. Zunehmend setzt sich aber eine Verwendung durch, die lediglich unerwünschte Ungleichbehandlungen erfasst.

Einzelne gesetzliche Regelungen verbieten Diskriminierungen aufgrund bestimmter Kriterien und in bestimmten Kontexten. In der oben (Kapitel 5.3) entwickelten Fehlertypologie sind sie

[114] So BeckOK Grundgesetz/Kischel, GG Art. 3 Rn. 184, bezogen auf die Diskriminierung wegen des Geschlechts.

den Fehlern der Entscheidungsgrundlage (Kapitel 5.3.4) oder gegebenenfalls auch deren Würdigung (Kapitel 5.3.5) zuzuordnen.

So verbietet Art. 3 Abs. 3 GG dem Staat eine Ungleichbehandlung aufgrund einer Reihe von Kriterien wie Geschlecht oder Abstammung. Zu dieser Norm liegt eine umfangreiche Rechtsprechung vor, insbesondere betreffend die Benachteiligung wegen des Geschlechts. Problemschwerpunkte liegen im Bereich der mittelbaren Benachteiligung (vgl. Kapitel 5.4.1) und der Rechtfertigungsgründe. Auch europarechtlich sind Gleichbehandlungsgrundsätze kodifiziert, etwa in Art. 18 AEUV bezüglich der Staatsangehörigkeit sowie in Art. 21, 23 der Charta der Grundrechte der Europäischen Union bezüglich einer Reihe von Merkmalen.

Diskriminierungen können auch wettbewerbsrechtlich problematisch sein: Die Konditionen- oder Preisdiskriminierung kann nach Art. 102 AEUV den Missbrauch einer marktbeherrschenden Stellung darstellen.[115] Auch § 42 Abs. 2 TKG enthält ein Diskriminierungsverbot, das den Wettbewerb sichern soll: Ein Unternehmen mit beträchtlicher Marktmacht darf anderen Unternehmen beim Zugang zu angebotenen Leistungen keine schlechteren Bedingungen einräumen als sich selbst bzw. den eigenen Tochter- oder Partnerunternehmen. Vorliegend sollen lediglich verbraucherrelevante Diskriminierungen und Diskriminierungsverbote betrachtet werden. Soweit die Rechtsprechung auf Prinzipien zurückgreift, die etwa im Kontext der verfassungsrechtlichen Diskussion entwickelt worden sind, werden diese unten genauer dargestellt.

Auch datenschutzrechtliche Normen werden zum Teil so verstanden, dass sie Diskriminierungen vorbeugen sollen.[116] Zunächst soll hier aber das Allgemeine Gleichbehandlungsgesetz betrachtet werden, das Diskriminierungen aufgrund von „Rasse oder wegen der ethnischen Herkunft, des Geschlechts, der Religion oder Weltanschauung, einer Behinderung, des Alters oder der sexuellen Identität“ (§ 1 AGG) verhindern soll und damit wesentliche Auswirkungen auch auf den Einsatz von ADM-System hat: Das Gesetz hat einen breiten Anwendungsbereich im Zivilrecht und betrifft in besonderem Ausmaß natürliche Personen in ihrer Rolle als Verbraucher.

Wie auch einige andere gesetzliche Regelungen knüpft das AGG begrifflich nicht an Diskriminierungen, sondern an Benachteiligungen an. Es sei aber darauf hingewiesen, dass – auch wenn der Begriff weit verstanden wird – „neutrale“ Diskriminierungen, bei denen unterschiedliche Gruppen unterschiedlich, aber gleichwertig behandelt werden, selten rechtlich relevant sind. Eine rechtlich relevante Ungleichbehandlung geht in der Regel mit Bevorzugungen und Benachteiligungen einher. Wird eine Gruppe bevorzugt, bedeutet dies immer gleichzeitig eine Benachteiligung anderer Gruppen[117]. Ansprüche aus dem AGG stehen aber aus offensichtlichen Gründen nur benachteiligten Personen zu.

Das AGG findet unter anderem bei einer Reihe von arbeitsrechtlichen Vorgängen Anwendung. Zwar soll das Arbeitsrecht an sich vorliegend nicht untersucht werden; in der entsprechenden Rechtsprechung werden aber auch Kriterien zur Erkennung einer Benachteiligung entwickelt, die wir im Folgenden wieder aufgreifen werden.

Daneben ist das AGG anwendbar bei Benachteiligungen betreffend „den Zugang zu und die Versorgung mit Gütern und Dienstleistungen, die der Öffentlichkeit zur Verfügung stehen, einschließlich von Wohnraum“ (§ 2 Abs. 1 Nr. 8 AGG). Tatsächlich unzulässig ist eine

[115] Grabitz et al./Deselaers, AEUV Art. 102 Rn. 420.
[116] BeckOK DatenschutzR/Schild, DS-GVO Art. 4 Rn. 188.
[117] So kann z.B. auch im Arbeitsrecht die Vorenthaltung eines Vorteils als Benachteiligung verstanden werden, vgl. ErfK/Preis, BGB § 612a Rn. 10.

Benachteiligung aus einem der genannten Gründe[118] nach § 19 Abs. 1 AGG aber nur bei der Begründung, Durchführung und Beendigung zivilrechtlicher Schuldverhältnisse, die

> *„typischerweise ohne Ansehen der Person zu vergleichbaren Bedingungen in einer Vielzahl von Fällen zustande kommen (Massengeschäfte) oder bei denen das Ansehen der Person nach der Art des Schuldverhältnisses eine nachrangige Bedeutung hat und die zu vergleichbaren Bedingungen in einer Vielzahl von Fällen zustande kommen“* (§ 19 Abs. 1 Nr. 1 AGG)

sowie bei solchen, die eine privatrechtliche Versicherung zum Gegenstand haben (§ 19 Abs. 1 Nr. 2 AGG). Lediglich eine Benachteiligung wegen der Rasse oder der ethnischen Herkunft ist auch bei der Begründung, Durchführung und Beendigung sonstiger zivilrechtlicher Schuldverhältnisse grundsätzlich unzulässig (§ 19 Abs. 2 AGG).

Bereits nach § 19 Abs. 1 AGG sind damit aber die Fälle, die den Schwerpunkt der vorliegenden Studie bilden, zumindest potenziell erfasst:

- Kreditscoring, soweit Kredite im Massengeschäft vergeben werden – wie dies bei Kleinkrediten wohl allgemein anzunehmen sein wird.[119] Auch wenn personenbezogene Daten über den Kreditnehmer erhoben werden, heißt das nicht, dass das Ansehen der Person eine wesentliche Bedeutung hat – zumindest nicht, sofern die Daten lediglich einer Schufa-Abfrage oder einer routinemäßigen Bonitätsprüfung dienen.[120]

- Preisdifferenzierung, da die Norm nicht nur darauf abstellt, ob ein Schuldverhältnis mit der betroffenen Person überhaupt eingegangen wird, sondern allgemein Benachteiligungen (also auch die Bedingungen des Geschäfts betreffend) bei der Begründung des Schuldverhältnisses erfasst.[121]

- Versicherungsrating, da Versicherungsgeschäfte in § 19 Abs. 1 Nr. 2 AGG ausdrücklich erwähnt sind.

Die genannten Massengeschäfte sind auch gerade diejenigen, bei denen eine algorithmische Entscheidungsfindung für die Anbieter besonders interessant ist.

Eine Benachteiligung kann auch in den genannten Fällen gerechtfertigt sein. § 20 AGG enthält eine Auflistung von Rechtfertigungsgründen; so ist beispielsweise eine Benachteiligung bei Versicherungen zulässig, „wenn diese auf anerkannten Prinzipien risikoadäquater Kalkulation beruht, insbesondere auf einer versicherungsmathematisch ermittelten Risikobewertung unter Heranziehung statistischer Erhebungen“ (§ 20 Abs. 2 AGG).

Es gibt keine Anhaltspunkte, wonach das AGG lediglich unmittelbar durch Menschen begründete Benachteiligungen erfassen soll. Eine Diskriminierung durch algorithmische

[118] Mit Ausnahme der Weltanschauung, die in § 19 Abs. 1 AGG nicht genannt wird.
[119] Die Gesetzesbegründung (BT-Drs. 16/1780, S. 42) geht zwar davon aus, dass Kreditgeschäfte „meist auf einer individuellen Risikoprüfung“ beruhen und es sich deshalb „regelmäßig nicht um Massengeschäfte“ handelt; mit MüKoBGB/Thüsing, AGG § 19 Rn. 24, ist aber davon auszugehen, dass diese Frage je nach Typ des Kredits einzeln zu beantworten ist.
[120] Ernst et al. 2013, AGG § 19 Rn. 4.
[121] Eine ausführliche Darstellung der Preisdifferenzierung nach Geschlecht findet sich bei Heiden/Wersig 2017.

Entscheidungen ist ohne weiteres denkbar[122]; sie lässt sich lediglich, falls dies gewünscht wird, beim Einsatz eines Algorithmus einfacher vermeiden als bei menschlichen Entscheidern.

Fraglich ist nun aber einerseits, wie weit der Begriff der Benachteiligung überhaupt reicht, und andererseits, wie sich eine Benachteiligung erkennen lässt.

5.4.1 Unmittelbare Benachteiligungen

Sogenannte unmittelbare Benachteiligungen sind vergleichsweise unproblematisch[123]. Sie sind in § 3 Abs. 1 AGG definiert:

„Eine unmittelbare Benachteiligung liegt 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."

Hier geht es also schlicht darum, ob die Anwendung eines bestimmten Kriteriums aus dem genannten Katalog zu einer Ungleichbehandlung führt. Bei einer algorithmischen Entscheidungsfindung lässt sich dies, sofern die Möglichkeit zum Testen des Algorithmus überhaupt besteht, einfach überprüfen. Dazu wird der Algorithmus bei ansonsten unveränderter Datengrundlage mit einem veränderten Wert des zu prüfenden Kriteriums erneut angewandt (vgl. dazu auch oben Kapitel 4.2.3 und 4.4.3.2/3). In der Praxis von durch Menschen getroffenen Entscheidungen kann die Frage nach der Vergleichbarkeit einer Situation schwierig zu beantworten sein.[124] Es ist aber zumindest ein Fall bekannt geworden, in dem eine unmittelbare Benachteiligung durch einen Menschen mit Hilfe eines Testverfahrens nachgewiesen wurde: Nachdem eine Mietinteressentin mit türkischem Namen auf ihre E-Mails hin nicht zu einem Besichtigungstermin für zwei Wohnungen eingeladen worden war, schickte ein Zeuge jeweils mehrere Anfragen mit erfundenen türkisch bzw. deutsch klingenden Namen als Absender an den Vermieter. Trotz ansonsten identischer Angaben wurden die fiktiven Interessenten mit deutsch klingenden Namen jeweils eingeladen, diejenigen mit türkisch klingenden Namen nicht. Das Amtsgericht Hamburg-Barmbek sah hierin ein hinreichendes Indiz für eine Benachteiligung aufgrund der ethnischen Herkunft.[125]

Bei algorithmischen Entscheidungen lassen sich einzelne Merkmale ohne weiteres abändern, um deren Einfluss auf eine Entscheidung zu untersuchen. Somit besteht für entsprechende Testverfahren mehr Spielraum als bei menschlichen Entscheidungen. Hacker[126] argumentiert, eine unmittelbare Diskriminierung liege nur bei bewusster oder unbewusster Voreingenommenheit (*explicit/implicit bias*) des Entscheiders vor, der im Lernprozess des maschinellen Lernverfahrens Klassifizierungen vorgibt oder bewusst unausgewogene Trainingsdaten verwendet (wie in Kapitel 4.2.1 beschrieben). Folgt man dieser Argumentation, bestünde bei algorithmischen Entscheidungen nur noch wenig Raum für unmittelbare Benachteiligungen ohne subjektives Vorsatzelement. Für die hier dargestellten Fälle, in denen ein Test eine unterschiedliche Entscheidung aufgrund eines

[122] So auch Dzida/Groh, NJW 2018, 1917, 1918.
[123] Dzida/Groh, NJW 2018, 1917, 1919, weisen allerdings darauf hin, dass die Voraussetzungen einer unmittelbaren Benachteiligung durch Algorithmen zumindest bei der Arbeitnehmerauswahl in der Praxis nur selten erfüllt sein dürften.
[124] Vgl. BeckOK ArbR/Roloff, AGG § 3 Rn. 7-10.
[125] AG Hamburg-Barmbek, Urteil vom 3. Februar 2017 – 811b C 273/15 –, juris.
[126] Hacker 2018, Teaching Fairness to Artificial Intelligence, Working Paper, April 2018, S. 9 f.

geschützten Attributs nachweist, ist die Kausalität – also die Benachteiligung „wegen“ dieses Attributs im Sinne von § 3 Abs. 1 AGG – unseres Erachtens aber gegeben.

Sofern keine entsprechende Testmöglichkeit besteht, wird der Nachweis unmittelbarer Benachteiligungen schwieriger. Hier kann gegebenenfalls auf statistische Methoden zurückgegriffen werden, wie sie im folgenden Kapitel – bezogen auf mittelbare Benachteiligungen – dargestellt werden.

5.4.2 Mittelbare Benachteiligungen

Neben den unmittelbaren erfasst das AGG auch sogenannte mittelbare Benachteiligungen, die in § 3 Abs. 2 AGG definiert sind:

„Eine mittelbare Benachteiligung liegt vor, wenn dem Anschein nach neutrale Vorschriften, Kriterien oder Verfahren Personen wegen eines in § 1 genannten Grundes gegenüber anderen Personen in besonderer Weise benachteiligen können, es sei denn, die betreffenden Vorschriften, Kriterien oder Verfahren sind durch ein rechtmäßiges Ziel sachlich gerechtfertigt und die Mittel sind zur Erreichung dieses Ziels angemessen und erforderlich.“

Obwohl der Wortlaut dies nahelegt, reicht die bloße Gefahr, dass eine Benachteiligung eintreten könnte, nicht aus, um eine mittelbare Benachteiligung anzunehmen.[127] Die Norm verlangt aber nur, dass die Benachteiligung Folge (und nicht zwingend Ziel) der jeweiligen Vorschrift, des Kriteriums oder des Verfahrens ist. Dennoch ist ihr Ziel, die Diskriminierung aufgrund eines der in § 1 AGG genannten Kriterien über einen Umweg bzw. mit Hilfe eines Vorwands zu verhindern.[128] Wir haben in Kapitel 4.2.2 dargelegt, dass die Gefahr einer solchen Diskriminierung auch bei der algorithmischen Entscheidungsfindung besteht, selbst wenn „geschützte“ Attribute wie das Geschlecht dem Algorithmus nicht direkt zugänglich sind.

Die Reichweite des Verbots mittelbarer Benachteiligungen lässt sich im Gegensatz zu unmittelbaren Benachteiligungen nur schwierig bestimmen. Hinzu kommt ein Nachweisproblem, das zwar grundsätzlich auch bei unmittelbaren Benachteiligungen auftreten kann – sofern kein Einblick in das Entscheidungsverfahren und keine Testmöglichkeit besteht. In der Praxis stellt sich die Erkennbarkeit einer mittelbaren Benachteiligung aber als deutlich schwieriger heraus, so dass die Problematik an dieser Stelle erläutert werden soll (vgl. hierzu auch das Beispiel in Kapitel 4.2.2).

Im Grundsatz erfordert eine mittelbare Benachteiligung „eine prozentual wesentlich stärkere Belastung einer Gruppe“[129]. Eine statistische Betrachtungsweise wird mit dieser Annahme nahegelegt. Hiermit stellt sich einerseits, wie bereits bei der unmittelbaren Diskriminierung, die Frage nach der Wahl der Vergleichsgruppe. Andererseits ist zu klären, wie stark der Zusammenhang (etwa als Korrelation ausgedrückt) zwischen dem Kriterium aus § 1 AGG und dem Ergebnis, das sich bei der Entscheidung ergibt, sein muss.

Bei der Wahl der Vergleichsgruppe scheinen in den bisher durch das BAG und den EuGH entschiedenen Fällen lediglich einzelne Kriterien relevant gewesen zu sein. Wie schon im

[127] BeckOK ArbR/Roloff, AGG § 3 Rn. 17; MüKoBGB/Thüsing, AGG § 3 Rn. 41, allerdings mit dem Hinweis auf die Gesetzesbegründung (BT-Drs. 17/1780, S. 33), die dem entgegenstehend bereits eine konkrete Gefährdung ausreichen lassen will.
[128] MüKoBGB/Thüsing, AGG § 3 Rn. 24.
[129] MüKoBGB/Thüsing, AGG § 3 Rn. 31.

Fall der unmittelbaren Diskriminierung dürften die praktischen Schwierigkeiten, die sich bei durch Menschen getroffenen Entscheidungen ergeben, bei algorithmischer Entscheidungsfindung zunächst zurücktreten: Sofern nur einzelne Kriterien betrachtet werden, lässt sich leicht simulieren, ob die Entscheidung bei Nichtberücksichtigung des Kriteriums anders ausgefallen wäre. Hier wird also auf die statistische Betrachtung der Gruppe verzichtet und stattdessen lediglich auf den Einzelfall abgestellt.

Der Zusammenhang zwischen dem Kriterium aus § 1 AGG und der getroffenen Entscheidung wird in der Rechtsprechung im Wesentlichen auf den Zusammenhang des Kriteriums aus § 1 AGG mit dem entscheidungsrelevanten Kriterium zurückgeführt: So wird berichtet, im Gros der Fälle sei es um Diskriminierung von Teilzeitbeschäftigten gegangen, unter denen im Durchschnitt über 90 % Frauen seien.[130] Eine scharfe Definition findet sich in der Literatur nicht, wohl auch wegen dieses in vielen Fällen starken Zusammenhangs beider Kriterien[131]. Auch in der verfassungsrechtlichen Literatur findet sich keine scharfe Grenze; das Bundesverfassungsgericht hat aber jedenfalls einen Frauenanteil von über 75 % bei den teilzeitbeschäftigten Beamten und Richtern ausreichen lassen, um eine Regelung, die diese Gruppe benachteiligt, als wegen des Geschlechts benachteiligend im Sinne des Art. 3 Abs. 3 GG anzusehen.[132] In der Literatur findet sich der gleiche Prozentsatz auch in einem anderen Kontext: Wenn statistisch signifikant nachgewiesen sei, dass die Wahrscheinlichkeit einer positiven Entscheidung für ein Mitglied einer geschützten Gruppe höchstens bei 75 % der Wahrscheinlichkeit einer positiven Entscheidung für ein Mitglied der (privilegierten) Vergleichsgruppe liegt, sei eine Benachteiligung gegeben.[133]

Auch in Entscheidungen des BAG und des EuGH werden statistische Vergleiche der betrachteten (sich in einem Kriterium aus § 1 AGG unterscheidenden) Gruppen herangezogen.[134] Es ist also davon auszugehen, dass auch Konzepte aus der Statistik, wie etwa die statistische Signifikanz des Unterschieds zwischen Gruppen, heranzuziehen sind. Diese werden in der vorliegenden Rechtsprechung aber nicht durchgehend erörtert; so hat der EuGH bereits entschieden, dass eine schlechtere Vergütung für Mehrarbeit von Teilzeitkräften bereits dann gegen das Gebot der Entgeltgleichheit verstößt, wenn die entsprechende Regelung „tatsächlich prozentual erheblich mehr Frauen als Männer benachteiligt“[135] (sofern keine Rechtfertigung vorliegt); wie ein „Erheblichkeitskriterium“ aussehen könnte, wird aber dort nicht erörtert. Das BAG hingegen hat in einer Entscheidung[136] ausdrücklich auf die mangelnde statistische Signifikanz eines Unterschieds abgestellt, ohne jedoch ein bestimmtes Signifikanzniveau zu fordern oder die Wahrscheinlichkeit für das zufällige Zustandekommen des betrachteten Unterschieds zu errechnen. Die quantitativen Diskriminierungsbegriffe aus der Informatikforschung (vgl. Kapitel 4.3.2) können hier einen Anhaltspunkt geben; der Rechtsprechung lässt sich aber keine Festlegung auf einen einzelnen dieser Begriffe – und erst recht kein verlässlicher Schwellwert, ab dem von einer Diskriminierung auszugehen ist – entnehmen.[137]

[130] MüKoBGB/Thüsing, AGG § 3 Rn. 31; ErfK/Schlachter, AGG § 3 Rn. 10.
[131] Also des an sich zulässigen und des nach § 1 AGG unzulässigen Kriteriums.
[132] BVerfGE 121, 241 (256). Unter den vollzeitbeschäftigten Beamten und Richtern lag der Frauenanteil zum Vergleichszeitpunkt bei 29 %; dem Wortlaut der Entscheidung nach kam es darauf aber nicht an.
[133] Hacker 2018, Teaching Fairness to Artificial Intelligence, Working Paper, April 2018, S. 10.
[134] BeckOK BGB/Fuchs, AGG § 3 Rn. 6.
[135] EuGH, Urteil vom 6. Dezember 2007, Rs. C-300/06, Slg. 2007, I-10573 – Voß.
[136] BAG, Urteil vom 27. Januar 2011 – 8 AZR 483/09 –, juris, Rn. 30.
[137] In der US-amerikanischen Rechtsprechung gibt es jedoch eine ausführliche Erörterung der Problematik. In der Memorandum Order des District Judge Higginbotham im Fall Vuyanich v.

Nach der Literatur ist der statistische Nachweis nicht in allen Fällen nötig. Wann genau er entfallen kann, ist nicht in allen Einzelheiten geklärt. Nach Thüsing ist aber davon auszugehen, dass er zumindest dann entbehrlich ist, wenn ein Kriterium „typischerweise zur Benachteiligung wegen eines der Merkmale des § 1 geeignet ist“[138].

Es sei darauf hingewiesen, dass eine nach den bisher erörterten Grundsätzen festgestellte Ungleichbehandlung durchaus sachlich gerechtfertigt und daher nicht als Benachteiligung im Sinne des § 3 Abs. 2 AGG anzusehen sein kann. Eine Regelung, die überwiegend Frauen betrifft, kann also beispielsweise rechtmäßig sein, wenn ein sachlicher Grund besteht – beispielsweise physische Mindestanforderungen, die zur Erfüllung einer Aufgabe notwendig sind, aber von Frauen seltener erfüllt werden. Das Vorliegen eines solchen Grundes lässt sich offensichtlich nicht mit Methoden der Statistik oder Informatik prüfen.

5.4.3 Offene Fragen der Diskriminierung durch Algorithmen nach dem AGG

5.4.3.1 Bestimmung der Vergleichsgruppen und Untergruppen

Für die Zwecke der vorliegenden Studie bleiben allerdings noch Fragen offen, die die Definition der Diskriminierung bzw. Benachteiligung betreffen. Dies liegt im Wesentlichen daran, dass sich die Rechtsprechung und die Literatur bisher zumindest weit überwiegend mit der Frage befasst haben, inwieweit das Heranziehen bestimmter einzelner Kriterien, für die dies aufgrund einer sachverständigen, inhaltlichen Betrachtung naheliegt, zu einer ungerechtfertigten Benachteiligung führt. Im Bereich algorithmischer Entscheidungsfindung könnten sich folgende neue Probleme ergeben (vgl. dazu auch Kapitel 4.4.3.2):

- Die Anzahl herangezogener Kriterien für eine Entscheidung steigt, da sowohl das geschützte Attribut als auch Proxyvariablen zu prüfen sind. Dies stellt unter Umständen höhere Anforderungen (Menge und Repräsentativität) an den Testdatensatz, um auch kleinere Untergruppen (z.B. Menschen mit Behinderungen) statistisch verlässlich untersuchen zu können. Insbesondere bei kleineren Datensätzen ergibt sich das Risiko einer „zufälligen“ Korrelation, die Lernverfahren negativ beeinflussen kann. Will man jedes einzelne Entscheidungskriterium statistisch darauf testen, ob es mit einem der Kriterien aus § 1 AGG korreliert, kann sich je nach Qualität der Daten die Wahrscheinlichkeit eines fehlerhaft (zufällig) erkannten Zusammenhangs erhöhen. Folglich muss dann die betrachtete Grundgesamtheit der Daten geprüft und/oder ausgeweitet werden, um solche Fehler auszuschließen.

  Beispiel: Ein einfaches ADM-System entscheidet über die Vergabe von Krediten in einer kleinen Bankfiliale. Das System trifft 100 Entscheidungen, darunter 30 Ablehnungen. Es wird nun geprüft, ob ein geschütztes Attribut unter den abgelehnten Kandidaten häufiger auftritt, als dies aufgrund der gesamten Häufigkeit des Attributs zu erwarten wäre. Dies wird zunächst geprüft für ethnische Herkunft, Geschlecht,

Republic Nat. Bank of Dallas, 505 F. Supp. 224 (N.D. Tex. 1980) findet sich (in Abschnitt VI.E) eine Erörterung statistischer Indikatoren für eine Diskriminierung. Das Gericht legte jedoch gerade keinen einzelnen Indikator als rechtlich relevant fest, da es die Gefahr sah, Entscheider könnten ihre Entscheidungen mit Bezug auf den Indikator optimieren, ohne Diskriminierungen jedoch tatsächlich abzustellen.

[138] MüKoBGB/Thüsing, AGG § 3 Rn. 32.

Religion, Behinderungen und Alter; anschließend wird die Prüfung für 10 weitere Attribute durchgeführt, die mit den genannten korreliert sind. Mit hoher Wahrscheinlichkeit wird festgestellt, dass mindestens eines der Attribute Einfluss auf die Kreditvergabe hat, denn es ist sehr unwahrscheinlich, dass eine Gruppe von lediglich 30 Personen in Bezug auf jedes dieser 15 Attribute dem Durchschnitt der Gesamtbevölkerung entspricht.

- Darüber hinaus sollten auch Kriterienkombinationen untersucht werden – etwa ob ein Entscheidungsverfahren besonders muslimische Frauen über 50 Jahren benachteiligt; damit verschärft sich das genannte Problem. Solche Kriterien bzw. Kriterienkombinationen und ihre Gewichtung können gegebenenfalls auch durch Eingreifen in einen algorithmischen Lernprozess bewusst so gesteuert werden, dass sie zu einer mittelbaren Benachteiligung führen. Andererseits kann auch umgekehrt der Lernprozess gerade mit dem Ziel der Vermeidung von Benachteiligungen beeinflusst werden.

Die genannten Schwierigkeiten können sich in verschiedener Weise auswirken. So besteht einerseits ein Risiko, dass der Nachweis einer tatsächlich stattfindenden Diskriminierung nicht gelingt, da bei kleinen Fallzahlen kein ausreichendes Signifikanzniveau des Nachweises erreicht wird. Andererseits[139] besteht die Gefahr, dass fälschlicherweise angenommen wird, eine Diskriminierung sei statistisch nachgewiesen. Zwar weist Ernst[140] darauf hin, dass bei algorithmischen Entscheidungen (im Gegensatz zu menschlichen Entscheidungen) ausschließlich die im Algorithmus abgebildeten Kriterien eine Rolle spielten und daher das Indiz eines statistischen Nachweises regelmäßig die Beweislastregelung des § 22 AGG auslöse; ein korrekter statistischer Nachweis dürfte sich jedoch aus den genannten Gründen in der Praxis oft nicht führen lassen.

Wollte man diese Gefahren weitgehend ausschließen, wäre ein Ansatz die Festlegung einer Positivliste von Attributen, die durch ADM berücksichtigt werden dürfen – nämlich genau solche Attribute, von denen im Vorhinein bekannt ist, dass sie für die Entscheidung relevant sind und ihre Verwendung sachlich gerechtfertigt ist. Ein solcher Ansatz würde aber die Vorteile der algorithmischen Entscheidungsfindung – zu denen auch das Erkennen vorher unbekannter Zusammenhänge gehört – drastisch reduzieren.

Zusammenfassend lässt sich festhalten, dass es keinen statistisch korrekten und juristisch anerkannten Standard gibt, der sowohl mittelbare Diskriminierungen nachweisbar ausschließt als auch die Vorteile algorithmischer Entscheidungsfindung nutzt. Daher besteht weiterer interdisziplinärer Forschungsbedarf, wobei auch organisatorische Maßnahmen (z.B. Prüfung durch Experten) als Lösungsansatz zu untersuchen wären.

5.4.3.2 Herausforderungen der Anwendung quantitativer Fairnessbegriffe

Sobald die zu vergleichenden Gruppen bestimmt sind, bieten quantitative Fairnessbegriffe (siehe Kapitel 4.3.2) zahlreiche Möglichkeiten zur Ermittlung und Charakterisierung einer mittelbaren Benachteiligung mit Konzepten zur Vergleichbarkeit anhand der Genauigkeit und des spezifischen Fehlerverhaltens eines ADM-Modells. Es läge daher nahe, dass der Gesetzgeber und/oder die Rechtsprechung diese Konzepte aufgreifen und statistische Verfahren in Zukunft als Beweismittel und/oder Prüfungsmaßstab stärker Verwendung

[139] Insbesondere vor dem Hintergrund, dass Statistik nicht immer intuitiv nachvollziehbar ist.
[140] Ernst, JZ 2017, 1026, 1033.

finden. Im Idealfall würde dies zu größerer Transparenz der Anforderungen für ADM-Systeme führen und somit Rechtssicherheit und klare Auditverfahren für Verwender schaffen. Die Vorteile einer solch eleganten Lösung liegen auf der Hand, jedoch stehen einer praktischen Umsetzung dieser Musterlösung diverse Hindernisse im Weg, die wir im Folgenden kurz erörtern.

- Es ist juristisch zu klären, wie quantitative Fairnessbegriffe in Rechtsetzung und Prüfung einfließen können. Wie in den Kapiteln 4.3.2 und 5.4.3.1 mit Hinweisen auf weitere Literatur bereits angedeutet wurde, ist universelle Gleichheit zwischen zwei Gruppen nach allen Metriken in realistischen Szenarien regelmäßig unmöglich. Stattdessen ist in konkreten Anwendungsfällen zu ermitteln, welche Metriken unter den gegebenen Umständen zwischen den gleich zu behandelnden Gruppen am wichtigsten sind. Ob und wie eine solche Abwägung rechtlich gefordert, sachlich geprüft und juristisch sichergestellt werden kann, ist völlig offen und muss weiter erforscht werden.

- Es könnten sich in dieser Untersuchung auch neue oder veränderte Anforderungen an Rechtfertigungstatbestände ergeben. Wenn die mittelbare Benachteiligung durch ein ADM-System durch bestimmte statistische Fehlermuster zwischen zwei Gruppen gekennzeichnet ist, stellt sich die Frage, inwieweit Rechtfertigungen diese Fehlermuster aufgreifen können bzw. sollten.

- In welchen prozessualen Mitteln sind quantitative Gleichbehandlungsbegriffe relevant? Wie ist die Beweisführung für die mittelbare Diskriminierung in diesem Fall zu gestalten? Ist es individuellen Prozessparteien außerhalb einer institutionellen Aufsicht überhaupt möglich, statistische Analysen zu erstellen und anzuführen?

- Wie an verschiedenen Stellen dieser Studie angedeutet, ergeben sich in der Umsetzung Fragen über die Verfügbarkeit und Wartung von Testdatensätzen zur Prüfung von ADM-Systemen. Ist es erstrebenswert, eine Institution der Bankenaufsicht die Praxis der Kreditvergabe an Verbraucher regelmäßig durch Testdaten anhand rein quantitativer Kriterien prüfen zu lassen? Wenn ja, wie und von wem sollen diese Daten erstellt und gewartet werden? Oder ist ein individueller Testdatensatz wie im Fair Lending Audit der USA sinnvoller? Wie können sich betroffene Unternehmen darauf vorbereiten und intern quantitative Metriken zum Testen ihrer ADM-Systeme nutzen? Welche Rolle spielen hierbei potenzielle strukturelle Inkompatibilitäten zwischen Auditdatensätzen und den internen Datensätzen der Bank?

- Im Wege der Aufsicht festgestellte statistisch signifikante Ungleichbehandlungen durch eine Bank nach quantitativen Metriken sind im Regelfall noch nicht ausreichend, um auf einen Rechtsverstoß zu schließen. Dennoch gilt es einen gesamtheitlichen Auditprozess juristisch auszugestalten und quantitative Metriken sachdienlich und effektiv darin einzubetten. Im Fair Lending Audit der USA dienen statistisch ermittelte Ungleichheiten weitestgehend der Erkundung der Daten und der Gesamtprüfung der Vergabepraxis der Bank nach Einbeziehung von möglichen Rechtfertigungen (siehe Kapitel 3.4.4). Dies kann als Leitbeispiel für die Nutzung statistischer Mittel in Auditprozessen dienen, jedoch sollte davon ausgehend die Entwicklung eines vergleichbaren Rechtsinstituts in der deutschen Rechtsordnung unter Einbeziehung der relevanten technischen Aspekte gründlich erörtert werden.

- Wie kann vermieden werden, dass Unternehmen durch technische Finesse in der Modellentwicklung das statistische Audit zwar bestehen, aber ihre ADM-Systeme auf den eigenen Daten trotzdem unausgewogene Entscheidungen produzieren („gaming the system“)? Das Problem stellt sich in besonderer Weise, wenn nur kleine Gruppengrößen betrachtet werden oder die Diskriminierung aufgrund mehrerer Attribute untersucht werden soll (vgl. Kapitel 5.4.3.1).

- Gegebenenfalls fallen Anwender und Entwickler des algorithmischen Entscheidungsverfahrens auseinander; beispielsweise könnte ein Verfahren zur Auswahl von Arbeitnehmern von Arbeitgebern aus verschiedenen Branchen genutzt werden, so dass sich die Frage nach der zu betrachtenden Grundgesamtheit bei statistischen Vergleichen neu stellt (siehe auch Kapitel 4.4.2).

Zusammenfassend lässt sich sagen dass quantitative Fairness- und Gleichbehandlungsbegriffe das konzeptionelle Vokabular zur Prüfung von mittelbarer Benachteiligung zwar bereichern, jedoch bezüglich ihrer juristischen Verwertung und effektiven praktischen Anwendung als Prüfinstrumente noch erheblicher Forschungsbedarf besteht.

5.4.4 Zwischenfazit zur Diskriminierung durch algorithmische Entscheidungen

Algorithmische Entscheidungen können, ebenso wie durch Menschen getroffene Entscheidungen, diskriminierend sein. Sie haben zunächst den Vorteil, dass sie per se nicht emotional gefärbt sind; jedoch können sie aufgrund eines Fehlers oder aufgrund bewusster Auswahl von Parametern und Attributen durchaus zu rechtswidrigen Ergebnissen führen. Der Nachweis ist im Einzelfall bei unmittelbarer Benachteiligung, falls eine Testmöglichkeit besteht, einfach zu führen; alternativ können auch statistische Verfahren herangezogen werden (makroskopische Betrachtung), wenn die dafür benötigten Daten vorliegen und eine qualifizierte Vergleichbarkeitsmetrik bestimmt wurde. Bei mittelbarer Benachteiligung hilft ein Test im Einzelfall in der Regel nicht weiter, und die Kriterien für einen statistischen Vergleich lassen sich weniger einfach formulieren. Der Einzelne, der von einer AGG-widrigen benachteiligenden Entscheidung betroffen ist, hat die Möglichkeit, dagegen vorzugehen; eine Ex-ante-Prüfung ist jedoch nicht vorgesehen.

Es sei zudem darauf hingewiesen, dass in Literatur und Rechtsprechung für den Nachweis indirekter Benachteiligung zwar statistische Vergleiche herangezogen werden; es gibt aber keine einheitliche Linie, worauf genau dieser Vergleich angewendet wird (und welches Signifikanzniveau gegebenenfalls zu verwenden ist). Außerdem besteht noch kein Schulterschluss zwischen Metriken der Gleichbehandlung aus der Literatur im Bereich *Fair Machine Learning* einerseits und juristischen Konzepten in Literatur, Rechtsetzung sowie Rechtsprechung andererseits. Ein einheitliches Kriterium, anhand dessen algorithmische Entscheidungen überprüft werden können, lässt sich also derzeit nicht verlässlich angeben. Aus unserer Sicht besteht hier noch erheblicher Forschungsbedarf, um die in beiden Gebieten nötigen Entwicklungen zusammenzuführen und produktiv zu bündeln.

Mögliche Rechtfertigungsgründe für die Diskriminierung sind einer rein mathematischen Prüfung nicht zugänglich und erst im nächsten Schritt zu prüfen.

5.5 Regulierung algorithmischer Entscheidungen im Datenschutzrecht

Die zentrale Norm, die maschinelle Entscheidungen im deutschen Recht reguliert, war nach bisherigem Recht § 6a BDSG – mit Anwendbarkeit der Datenschutz-Grundverordnung (DSGVO) wurde er durch deren Art. 22 abgelöst. Wir betrachten an dieser Stelle ausschließlich die neue Rechtslage.

Zunächst ist festzustellen, dass die DSGVO insgesamt nur dann anwendbar ist, wenn *personenbezogene* Daten verarbeitet werden (vgl. Art. 2 Abs. 1 DSGVO). Der Begriff ist aber sehr weit gefasst; er beinhaltet „alle Informationen, die sich auf eine identifizierte oder identifizierbare natürliche Person […] beziehen" (Art. 4 Abs. 1 DSGVO). Auch der Begriff der Identifizierbarkeit wird weit ausgelegt; es genügt eine indirekte Zuordnungsmöglichkeit. Die Kenntnis des Namens der Person ist nicht erforderlich. Andererseits sind nicht alle nur irgendwie theoretisch denkbaren Möglichkeiten der Zuordnung von Daten zu einer Person zu berücksichtigen, sondern nur solche, „die von dem Verantwortlichen oder einer anderen Person nach allgemeinem Ermessen wahrscheinlich genutzt werden, um die natürliche Person direkt oder indirekt zu identifizieren, wie beispielsweise das Aussondern" (EG 26 Satz 3 DSGVO).

In der Praxis dürfte die Einschränkung der Anwendbarkeit auf die Verarbeitung personenbezogener Daten kaum Auswirkungen auf die Bedeutung des Art. 22 DSGVO haben. Aggregierte Daten über hinreichend große Gruppen (z.B. alle Bewohner eines Stadtviertels) sind per se zwar nicht personenbezogen; sie werden es aber, wenn sie mit einer Person in Verbindung gebracht werden. Die Aussage „Bewohner des Stadtteils X zwischen 30 und 40 Jahren haben bei Verbraucherkrediten ein erwartetes Ausfallrisiko von 2,3 %" ist also nicht personenbezogen; wird sie zusammengeführt mit der Aussage „Herr Schmidt ist ein 33-jähriger Bewohner des Stadtteils X", ist sie es. Sollen algorithmisch individuelle Entscheidungen getroffen werden, die natürliche Personen betreffen, ist also die Anwendbarkeit der DSGVO kaum zu umgehen.

Inhaltlich erfasst Art. 22 DSGVO zunächst alle Entscheidungen, die ausschließlich auf einer automatisierten Verarbeitung beruhen, sofern sie der betroffenen Person gegenüber eine rechtliche Wirkung entfalten oder diese in ähnlicher Weise erheblich beeinträchtigen. Solche Entscheidungen sind grundsätzlich verboten (Abs. 1 der Norm); die in Abs. 2 definierten Ausnahmen (Erlaubnistatbestände) haben aber eine große Reichweite. Sie sind verknüpft mit der Forderung nach angemessenen Schutzmaßnahmen (Abs. 2 lit. b und Abs. 3). Die Ausnahmen werden für besondere Kategorien personenbezogener Daten weiter eingeschränkt (Abs. 4). Wir betrachten die einzelnen Voraussetzungen in den folgenden Kapiteln.

5.5.1 Verbot automatisierter Entscheidungen

Nach Art. 22 Abs. 1 DSGVO hat der Betroffene das Recht, nicht einer ausschließlich auf einer automatisierten Verarbeitung beruhenden Entscheidung (folgend: „automatisierte Entscheidung") unterworfen zu werden, die ihn gegenüber rechtlicher Wirkung enthalte oder in ähnlicher Weise erheblich beeinträchtigt. Entgegen dem Wortlaut normiert Abs. 1 nach herrschender Meinung der Literatur nicht lediglich einen Anspruch des Betroffenen auf Unterlassung, sondern ein Verbot der automatisierten Entscheidung.[141]

[141] Kühling/Buchner-Buchner, Art. 22 Rn. 12; Martini in Paal/Pauly, Art. 22 Rn. 1, 29.

5.5.2 Begriff der automatisierten Entscheidung

Zunächst ist zu klären, was unter dem Begriff der automatisierten Entscheidung bzw. genauer der „ausschließlich auf einer automatisierten Verarbeitung – einschließlich Profiling – beruhenden Entscheidung“ (Art. 22 Abs. 1 DSGVO) zu verstehen ist.

Der Begriff des Profilings ist in Art. 4 Abs. 4 DSGVO legaldefiniert als „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“. Wird also die Bonität einer Person anhand personenbezogener Daten algorithmisch bewertet und nur anhand dessen über eine Darlehensvergabe entschieden, liegen klar ein Profiling und eine automatisierte Entscheidung vor. Auch Kreditscoring zählt zum Profiling.[142]

Es muss aber für eine automatisierte Entscheidung kein Profiling vorliegen – dieses Konzept ist lediglich „eingeschlossen“. Stattdessen ist zunächst jede Entscheidung erfasst, die auf einer automatisierten Verarbeitung personenbezogener Daten beruht; eine Einschränkung auf eine Bewertung der in der Definition des Profilings genannten Aspekte besteht nicht.[143]

Für die Zwecke der vorliegenden Studie ist der Begriff der „Entscheidung“ allerdings selbst noch genauer zu untersuchen. So sei zunächst auf die Abgrenzung zwischen einer (durch die Norm erfassten) Entscheidung und deren (nicht erfasster) Vorbereitung hingewiesen. Beispielsweise ist die Berechnung von Score-Werten durch die Schufa oder andere Auskunfteien noch keine Entscheidung (kann aber gegebenenfalls in eine algorithmische Entscheidung einfließen). Sofern ein Mensch mit Entscheidungskompetenz und Beurteilungsspielraum ausgestattet ist und die Letztentscheidung trifft, liegt zumindest keine „ausschließlich auf einer automatisierten Verarbeitung“ beruhende Entscheidung vor. Ein nachgeschaltetes reines „Abnicken“ reicht aber nicht aus.[144]

In der Literatur wird darauf hingewiesen, dass der Wortlaut der Norm gegebenenfalls einen sehr weiten Anwendungsbereich nahelegen könnte. Demnach wären „letztendlich triviale Wenn-dann-Entscheidungen wie Abhebungen am Geldausgabeautomaten“[145] erfasst. Dem ließe sich entgegenhalten, dass eine Entscheidung lediglich dann Art. 22 unterliegt, wenn sie der betroffenen Person „gegenüber rechtliche Wirkung entfaltet oder sie in ähnlicher Weise erheblich beeinträchtigt“; zudem ist selbst dann unter den Voraussetzungen der Absätze 2 und 3 die automatisierte Einzelentscheidung gegebenenfalls zulässig. Diese Einschränkungen helfen aber angesichts der Vielzahl der genannten „trivialen Wenn-dann-Entscheidungen“, deren sehr einfache Sachverhalte in einer menschlichen Überprüfung kaum anders beurteilt werden dürften, nicht weiter.

[142] Kühling/Buchner-Buchner, Art. 22 Rn. 22.
[143] So auch Schulz in Gola, DSGVO, 2017, Art. 22 Rn. 20; von Lewinski (in BeckOK Datenschutzrecht, DSGVO, Art. 22 Rn. 8; Wolff/Brink, 22. Auflage, Stand: 01.11.2017) sieht den Regelungsgegenstand auf die „Bewertung persönlicher Merkmale“ eingeschränkt.
[144] Schulz, Art. 22 Rn. 14-17; Ernst, JZ 2017, 1026, 1029 f. verweist auf die Vorgängerregelung im nationalen Recht (§ 6a Abs. 1 Satz 2 BDSG a.F.), wonach es darauf ankommt, ob eine inhaltliche Bewertung und darauf gestützte Entscheidung durch eine Person stattgefunden hat.
[145] Von Lewinski, Art. 22 Rn. 13.

Vielmehr ist davon auszugehen, dass ein Mindestmaß an Komplexität vorausgesetzt wird, bevor von einer tatsächlich automatisierten Entscheidung gesprochen werden kann.[146] Bei den genannten einfachen Entscheidungsregeln lässt sich die Entscheidung noch eher auf denjenigen zurückführen, der die Regel vorgegeben hat (Entscheidung „auf Vorrat").

Problematisch ist hierbei jedoch, dass für den Einzelnen nicht immer ohne weiteres ersichtlich ist, wie komplex das Verfahren ist, das zu einer bestimmten Entscheidung führt. So wird in der Literatur die Genehmigung von Kreditkartenverfügungen als Beispiel für die „trivialen Wenn-dann-Entscheidungen" genannt[147]; tatsächlich werden Kreditkartenverfügungen in der Regel aufgrund durchaus komplexer *Fraud-Detection*-Algorithmen genehmigt oder abgelehnt, die auf maschinellen Lernverfahren beruhen[148]. Es ist auch davon auszugehen, dass die Ablehnung einer Kreditkartenverfügung – die noch dazu gerade auf Auslandsreisen besonders wahrscheinlich ist, also dann, wenn auch wenige Alternativen zur Verfügung stehen – eine erhebliche Beeinträchtigung darstellt, so dass Art. 22 DSGVO in diesen Fällen eben doch Anwendung finden dürfte.

In den vorliegend zu betrachtenden verbraucherschutzrelevanten Fällen (und allgemein beim Einsatz maschineller Lernverfahren) dürfte in aller Regel von einer ausreichenden Komplexität des Entscheidungsfindungsverfahrens auszugehen sein, so dass die Anwendbarkeit des Art. 22 DSGVO aus diesem Gesichtspunkt heraus nicht in Frage stehen dürfte. Dies gilt auch für die vorliegend schwerpunktmäßig betrachteten Fälle, nämlich die Preisdifferenzierung[149] und die Kreditvergabe auf Grundlage maschineller Lernverfahren; reines Kreditscoring, das als Grundlage einer menschlichen Entscheidung dient, ist aus den oben genannten Gründen hingegen nicht erfasst (vgl. dazu auch Kapitel 5.5.4).

Nebenbei sei bemerkt, dass sich eine weitere Einschränkung aus Art. 22 Abs. 4 DSGVO ergibt; sie betrifft die Entscheidungsfindung basierend auf besonderen Kategorien personenbezogener Daten[150]. Diese ist nur auf Grundlage einer ausdrücklichen Einwilligung oder einer gesonderten Rechtsgrundlage und auch dann nur bei Einhaltung angemessener Schutzmaßnahmen zulässig.

5.5.3 Erlaubnistatbestände

Art. 22 Abs. 2 sieht drei Ausnahmen von dem Verbot aus Abs. 1 vor:

- Die automatisierte Einzelentscheidung ist zulässig, falls sie „für den Abschluss oder die Erfüllung eines Vertrags zwischen der betroffenen Person und dem Verantwortlichen erforderlich ist". Die Erforderlichkeit ist so zu verstehen, dass sie sich auf die *automatisierte* Entscheidung bezieht, nicht auf die Notwendigkeit,

[146] Von Lewinski, Art. 22 Rn. 12; Schulz, Art. 22 Rn. 20.
[147] Von Lewinski, Art. 22 Rn. 13.
[148] Ein Bericht über die praktischen Herausforderungen dieser Verfahren findet sich bei Dal Pozzolo/ Caelen/Le Borgne/Waterschoot/Bontempi, Learned lessons in credit card fraud detection from a practitioner perspective, in Expert Systems with Applications 41(10), Aug. 2014, S. 4915-4928.
[149] So mit Verweis auf eine erhebliche Beeinträchtigung aufgrund der finanziellen Wirkung auch Ernst, JZ 2017, 1026, 1034 f.
[150] Es handelt sich nach Art. 9 Abs. 1 DSGVO um „personenbezogene Daten, aus denen die rassische und ethnische Herkunft, politische Meinungen, religiöse oder weltanschauliche Überzeugungen oder die Gewerkschaftszugehörigkeit hervorgehen", sowie „genetische Daten, biometrische Daten zur eindeutigen Identifizierung einer natürlichen Person, Gesundheitsdaten oder Daten zum Sexualleben oder der sexuellen Orientierung einer natürlichen Person".

überhaupt eine Entscheidung zu treffen. Gründe könnten sowohl in der Geschwindigkeit als auch in der Anzahl zu treffender Entscheidungen liegen.[151]

- Sie kann außerdem aufgrund anderer Rechtsvorschriften der Europäischen Union oder des jeweiligen Mitgliedstaates zulässig sein.
- Schließlich ist die automatisierte Einzelentscheidung auch zulässig, wenn der Betroffene ausdrücklich eingewilligt hat; eine lediglich konkludente Einwilligung reicht also nicht aus. Auch sind die allgemeinen Anforderungen an eine datenschutzrechtliche Einwilligung zu beachten.[152]

Bereits der erste Ausnahmetatbestand dürfte in einer Vielzahl praktisch relevanter Fallkonstellationen erfüllt sein; die Schutzwirkung des Art. 22 entfällt dennoch nicht, denn in allen drei Fällen sind Schutzmaßnahmen für den Betroffenen vorzusehen. Erlaubt eine Rechtsvorschrift außerhalb der DSGVO die automatisierte Einzelentscheidung, so muss sie „angemessene Maßnahmen zur Wahrung der Rechte und Freiheiten sowie der berechtigten Interessen der betroffenen Person enthalten“ (Art. 22 Abs. 2 lit. b DSGVO).

Für die anderen Fälle gilt Abs. 3, wonach der Verantwortliche angemessene Maßnahmen zu treffen hat – darunter „mindestens das Recht auf Erwirkung des Eingreifens einer Person seitens des Verantwortlichen, auf Darlegung des eigenen Standpunkts und auf Anfechtung der Entscheidung“[153]. Dies führt – mit Ausnahme gegebenenfalls zukünftig möglicher Rechtsvorschriften, die die Öffnungsklausel des Art. 22 Abs. 2 lit. b nutzen – insgesamt dazu, dass die betroffene Person, sofern sie ihre Rechte geltend macht, nie gegen ihren Willen einer endgültigen automatisierten Einzelentscheidung unterworfen werden darf.

Weiter geht die Norm jedoch nicht; dies korrespondiert damit, dass sie für die automatisierte Entscheidungs*vorbereitung* gar nicht erst anwendbar ist. Sie kann nicht verhindern, dass sich der menschliche Entscheider überwiegend auf die Norm verlässt. Zwar wird gefordert, dass dieser eigenen Entscheidungsspielraum hat; dies zu widerlegen, kann in der Praxis aber problematisch sein.

Praxis: Wir führten ein Gespräch mit einem Vertreter der hessischen Datenschutzaufsicht (dem Hessischen Beauftragten für Datenschutz und Informationsfreiheit). Dort wird die Ansicht vertreten, dass Art. 22 DSGVO ein „stumpfes Schwert“ sei, da in der Praxis vollständig automatisierte Entscheidungen nur selten vorkämen. Lediglich die automatisierte Kreditvergabe bei Verbraucherkrediten für Unterhaltungselektronik sei eine relevante Ausnahme. Auch ein sehr geringer Spielraum seitens der Sachbearbeiter sei ausreichend, um aus dem Anwendungsbereich des Art. 22 herauszufallen.

5.5.4 Scoring im BDSG

Mit § 31 des neuen BDSG hat auch der nationale Gesetzgeber eine Regelung geschaffen, die Auswirkungen auf algorithmische Entscheidungen hat. Die Norm regelt – innerhalb des Anwendungsbereichs der DSGVO – in Abs. 1 die Zulässigkeit von Scoring („Verwendung eines Wahrscheinlichkeitswerts über ein bestimmtes zukünftiges Verhalten einer natürlichen Person zum Zweck der Entscheidung über die Begründung, Durchführung oder Beendigung eines Vertragsverhältnisses mit dieser Person“) allgemein, in Abs. 2 bezogen auf

[151] Von Lewinski, Art. 22 Rn. 43.
[152] Kühling/Buchner-Buchner, DS-GVO, Art. 22 Rn. 42.
[153] Gemeint ist die Einräumung des Rechts als Maßnahme.

Wahrscheinlichkeitswerte über die Zahlungsunfähigkeit und -unwilligkeit. Abs. 2 regelt dabei lediglich, welche Forderungen bei der Berechnung eines Wahrscheinlichkeitswerts durch Auskunfteien berücksichtigt werden können[154].

Ob § 31 BDSG mit der DSGVO vereinbar ist, ist nicht geklärt. Keine der Öffnungsklauseln der DSGVO erlaubt nach ihrem Wortlaut nationale Regelungen, die das Scoring einschränken.[155] Die Öffnungsklausel des Art. 22 Abs. 1 lit. b DSGVO bezieht sich lediglich auf Entscheidungen, die auf einer automatisierten Verarbeitung beruhen; Scoring wird aber nur als entscheidungsvorbereitende Maßnahme gesehen (und auch in der Praxis oft durch Auskunfteien durchgeführt, die selbst keine Entscheidungen treffen[156]). Bei einer weiten Auslegung des Art. 22 Abs. 1 lit. b DSGVO[157] oder wenn man § 31 BDSG nicht als datenschutzrechtliche Norm (sondern als „Schutzvorschrift gegen Diskriminierungen und Einschränkung der Privatautonomie“[158]) versteht, lässt sich die Norm aber halten.

Auch wenn Scoring per se lediglich rein entscheidungsvorbereitend ist, können Scoring-Verfahren natürlich bei der algorithmischen Entscheidungsfindung zum Einsatz kommen. Selbst wenn man die Anwendbarkeit von § 31 BDSG für das entscheidungsvorbereitende Scoring ablehnt, lässt sich die Norm doch unionsrechtskonform dahingehend auslegen, dass automatisierte Entscheidungen auf Basis von Scoring erfasst werden. Offen ist das Verhältnis zu den weiteren Erlaubnistatbeständen des Art. 22 Abs. 2 DSGVO; stützt man sich lediglich auf die Öffnungsklausel des Art. 22 Abs. 1 lit. b DSGVO, können diese Erlaubnistatbestände neben § 31 BDSG treten und kann die automatisierte Entscheidung trotz Verwendung von Scoring auf eine Einwilligung des Betroffenen gestützt werden.

Die Definition des Scorings setzt voraus, dass ein Wahrscheinlichkeitswert berechnet wird. Wahrscheinlichkeiten werden in der Mathematik üblicherweise so definiert, dass sie lediglich Werte im Intervall [0;1] annehmen können. Eine Beschränkung auf den mathematischen Wahrscheinlichkeitsbegriff würde jedoch dem Schutzzweck der Norm nicht gerecht. Auch grobe Klassifizierungen wie die Einteilung in eine von drei Risikoklassen oder Umrechnungen, etwa in eine Skala von 0 bis 20, sind also erfasst.

Scoring ist nach § 31 Abs. 1 BDSG nur zulässig, wenn vier Voraussetzungen erfüllt sind. Demnach müssen bei der vorherigen Verarbeitung der Daten die datenschutzrechtlichen Vorschriften eingehalten worden sein; es dürfen nicht ausschließlich Anschriftendaten für die Berechnung genutzt werden, und wenn Anschriftendaten für die Berechnung genutzt werden, muss der Betroffene hierüber unterrichtet werden. Schließlich müssen die zur Berechnung des Wahrscheinlichkeitswerts genutzten Daten auch „unter Zugrundelegung eines wissenschaftlich anerkannten mathematisch-statistischen Verfahrens nachweisbar für die Berechnung der Wahrscheinlichkeit des bestimmten Verhaltens erheblich“ sein. Wie diese Erheblichkeitsschwelle praktisch festgelegt werden kann, wird in der Literatur bislang aber nicht thematisiert.

[154] Eigentlich regelt Abs. 2 die Verwendung des berechneten Werts, hat aber natürlich eine direkte Auswirkung auf die Auskunfteien, die den Wert berechnen.
[155] Kühling et al. 2016, S. 440-445.
[156] Wohl aber kann der errechnete Score-Wert später durch eine Bank bei der Entscheidung über die Kreditvergabe, sei es durch einen Sachbearbeiter oder ein ADM-System, berücksichtigt werden. Gegebenenfalls (bei Verwendung eines ADM-Systems) findet Art. 22 DSGVO an dieser Stelle Anwendung.
[157] Dazu ausführlich Kühling et al. 2016, S. 440-445.
[158] Kühling NJW 2017, 1985, 1988.

Praxis: Wir führten ein Gespräch mit einem Vertreter der hessischen Datenschutzaufsicht (dem Hessischen Beauftragten für Datenschutz und Informationsfreiheit). In diesem Gespräch erfuhren wir, dass zur Prüfung, ob das Scoring der Schufa den Anforderungen der Vorgängernorm § 28b BDSG a.F. genügte, Gutachten vorgelegen haben. Daraus hätten sich keine Zweifel an der wissenschaftlichen Fundiertheit des Verfahrens und der Adäquanz der verwendeten Attribute ergeben; für einen besonders strengen Prüfungsmaßstab habe es keinen Anlass gegeben.

Fraglich ist auch, ob jedes einzelne Attribut für sich genommen erheblich sein muss oder auch die Erheblichkeit von Attributkombinationen ausreicht. Dies ist gerade für maschinelle Lernverfahren relevant, die ihre Stärke insbesondere dann ausspielen, wenn zahlreiche Attribute bekannt sind – wobei der Beitrag einzelner Attribute für die Entscheidung aber gegebenenfalls gering sein kann. Zudem ist dieser Beitrag in der Regel nicht im Vorhinein bekannt; bei einer engen Auslegung der Norm besteht die Möglichkeit einer zu starken Einschränkung der Anwendbarkeit maschineller Lernverfahren. Dies gilt mehr noch für innovative und zukünftige Anwendungen als für solche, die das momentan wohl am weitesten verbreitete statistische Modell der logistischen Regression nutzen, da dieses hinsichtlich der Erkennung nichtlinearer Zusammenhänge beschränkt ist (vgl. Kapitel 4.1).

Inwieweit eine Einschränkung auf einzelne, gegebenenfalls sogar im Vorhinein bekannte Attribute – die die Grundannahmen des Big-Data-Paradigmas betrifft – erwünscht ist, ist politisch und juristisch noch zu klären. Die Diskussion steht hier allerdings erst am Anfang.

Es sei darauf hingewiesen, dass der Bundesrat in seiner Stellungnahme[159] zum Datenschutz-Anpassungs- und -Umsetzungsgesetz EU deutlich weitergehende Regelungen zum Scoring gefordert hat, so z.B. eine „Eingrenzung der zulässigen Datenarten und -quellen bei Aufstellung von Scoring-Verfahren“. Diese sind allerdings nicht in das Gesetz eingeflossen.

5.5.5 Informationspflichten

Als weiteres Schutzinstrument bei automatisierten Entscheidungen enthält die DSGVO in Art. 13 und 14 spezifische Informationspflichten des Verantwortlichen. Art. 13 und 14 DSGVO regeln Informationen, die der betroffenen Person mitzuteilen sind, wenn die Daten direkt bei ihr erhoben werden (Art. 13) bzw. wenn sie nicht direkt bei ihr erhoben werden (Art. 14). In beiden Fällen ist über „das Bestehen einer automatisierten Entscheidungsfindung einschließlich Profiling gemäß Artikel 22 Absätze 1 und 4“ zu informieren. Dazu gehören „– zumindest in diesen Fällen [also jenen der Absätze 1 und 4] – aussagekräftige Informationen über die involvierte Logik sowie die Tragweite und die angestrebten Auswirkungen einer derartigen Verarbeitung für die betroffene Person“ (Art. 13 Abs. 2 lit. f und wortgleich Art. 14 Abs. 2 lit. g). Inwieweit solche Informationen zur Kenntnis genommen werden und ob – etwa im bereits genannten Beispiel der Autorisierung einer Kreditkartentransaktion – die betroffene Person eine einmal erhaltene Information mit einer gegebenenfalls Jahre später getroffenen Entscheidung in Verbindung bringt, ist allerdings offen.

Auch ist darauf hinzuweisen, dass der Anwendungsbereich der genannten Informationspflicht demjenigen des Art. 22 entspricht und somit auf automatisierte

[159] BR-Drs. 110/17 (Beschluss), S. 30-32.

Entscheidungen (nach dem vorherrschenden engen Verständnis dieser Norm also nicht auf die Entscheidungsvorbereitung[160]) begrenzt ist.[161]

Die genannten „aussagekräftigen Informationen über die involvierte Logik" beinhalten keine Offenlegung des verwendeten Algorithmus oder der verwendeten Trainingsdaten eines maschinellen Lernverfahrens.[162] Für die meisten betroffenen Personen wären diese nicht hilfreich und somit auch nicht aussagekräftig; Martini stellt auf die „grundsätzliche Entscheidungsstruktur, die den implementierten Algorithmen zugrunde liegt", ab. [163] Da die Trainingsdaten in vielen Fällen personenbezogene Daten Dritter sein werden[164], wäre ein bedingungsloser Anspruch auf deren Herausgabe ohnehin problematisch. Ein Anspruch auf eine genaue Begründung der einzelnen Entscheidung lässt sich auch dem Wortlaut der Norm nicht entnehmen. Mit Ernst ist davon auszugehen, dass die Informationspflicht die Nennung der für eine Entscheidung herangezogenen Kriterien umfasst – der Nutzen dieser Nennung für den Betroffenen ist aber begrenzt.[165]

5.5.6 Zwischenfazit zur datenschutzrechtlichen Algorithmenregulierung

Auf dem aktuellen Stand der Diskussion ist festzuhalten, dass Art. 22 DSGVO im Zusammenhang mit den Informationspflichten aus Art. 13 und 14 DSGVO bereits einen Großteil der hier interessierenden algorithmischen Entscheidungen – jedoch nicht die Entscheidungsvorbereitung[166] – erfasst. Im Kern sichert er die Möglichkeit einer menschlichen Überprüfung zu, sofern die Letztentscheidung nicht – wie bei der Verwendung von Scoring-Verfahren oft der Fall – ohnehin durch einen Menschen gefällt wird.

Es wird zu diskutieren sein, ob die getroffenen Regeln auch praktisch ausreichen, da eine Offenlegung von Entscheidungsgründen gerade nicht gefordert ist und auch die Zielrichtung des Art. 22 nicht die Fehlerminimierung des Entscheidungsverfahrens, sondern vielmehr nur den Umgang mit getroffenen, möglicherweise fehlerhaften Entscheidungen umfasst[167]. Wie auch im Fall der AGG-widrigen Diskriminierung ist im Fall des Art. 22 DSGVO zunächst keine Ex-ante-Prüfung vorgesehen. Bezogen auf alle drei Gefährdungsszenarien (aus Kapitel 2.2) ist festzuhalten, dass keines dieser Szenarien vollständig durch Artikel 22 unterbunden werden kann. Die *Richtigkeit der Entscheidungen* und der *Schutz vor Diskriminierung* spielen gegebenenfalls eine Rolle im Rahmen der Maßnahmen nach Abs. 3; auch das gegebenenfalls bestehende Recht auf menschliche Überprüfung der Entscheidung kann im Einzelfall hilfreich sein. Jenseits des Einzelfalls verhindern kann Art. 22 aber weder inhaltlich unrichtige noch diskriminierende Entscheidungen. Die an Art. 22 anknüpfenden Informationspflichten erhöhen die Transparenz über algorithmische Entscheidungen; ihr

[160] Kritisch zu diesem engen Anwendungsbereich Wischmeyer, AöR 2018, 1, 22 f.
[161] Vgl. Martini, JZ 2017, 1017, 1020.
[162] So auch, bezogen auf den Programmcode, Martini, JZ 2017, 1017, 1020. Auch Kühling/Buchner-Buchner, Art. 22 Rn. 35, sehen keinen Anspruch auf Offenlegung des Algorithmus. Bereits auf Grundlage des § 34 Abs. 4 S. 1 Nr. 4 BDSG a.F. hatte der BGH einen entsprechend weiten Anspruch auf Auskunft über die genaue Berechnungsformel für den Schufa-Score abgelehnt (BGH, Urteil vom 28.01.2014 – VI ZR 156/13 –, BGHZ 200, 38).
[163] Martini, JZ 2017, 1017, 1020.
[164] Auch beim trainierten Modell lässt sich nicht ausschließen, dass Rückschlüsse auf personenbezogene Daten Dritter gezogen werden können.
[165] Ernst, JZ 2017, 1026, 1033.
[166] Martini, JZ 2017, 1017, 1020.
[167] So weist auch Ernst, JZ 2017, 1026, 1030 darauf hin, dass ein Mensch dieselbe Entscheidung treffen kann, die als maschinelle Entscheidung unzulässig wäre – es also nicht auf den Inhalt der Entscheidung ankomme.

Anwendungsbereich geht aber nicht über den des Art. 22 hinaus, so dass auch hier die – in der Praxis äußerst relevante – Entscheidungsvorbereitung außer Acht bleibt.

Es sei jedoch darauf hingewiesen, dass algorithmische Entscheidungen lediglich ein Sonderfall der Verarbeitung (Definition in Art. 4 Nr. 2 DSGVO) sind; somit sind die allgemeinen datenschutzrechtlichen Regeln anwendbar. Dies gilt auch für Art. 35 DSGVO, der die Datenschutz-Folgenabschätzung regelt. Bei einem „hohen Risiko für die Rechte und Freiheiten natürlicher Personen“ (Art. 35 Abs. 1) durch die algorithmische Entscheidung muss der Verantwortliche die Verarbeitungsvorgänge, deren Notwendigkeit und Verhältnismäßigkeit gemäß dem verfolgten Zweck, die Risiken und Abwehrmaßnahmen dokumentieren. Es ist nicht davon auszugehen, dass algorithmische Entscheidungen grundsätzlich solch hohe Risiken mit sich bringen; es wäre aber zumindest denkbar, die Pflicht zur Durchführung einer Folgenabschätzung auszuweiten.

Offen ist die genaue Bedeutung des § 31 BDSG für ADM-Systeme. Dies liegt einerseits an der noch nicht abschließend geklärten Frage nach seiner Vereinbarkeit mit der DSGVO; andererseits ist die Anforderung, dass die verwendeten Daten für die Berechnung der Wahrscheinlichkeit des bestimmten Verhaltens „erheblich“ sein müssen, auslegungsbedürftig.

Hacker[168] führt an, dass das Datenschutzrecht im Zusammenspiel mit dem Antidiskriminierungsrecht geeignet sein könne, um die Fairness von ADM herzustellen. Er verweist dazu auf das Auskunftsrecht nach Art. 15 Abs. 1 DSGVO, die Befugnisse der Aufsichtsbehörden, Datenschutzüberprüfungen vorzunehmen, sowie die genannte Datenschutz-Folgenabschätzung. Ob die Reichweite der Normen ausreicht, um dieses Ziel zu erreichen, ist jedoch unklar; hierzu bedarf es noch weiterer Forschung.

5.6 Regulierung algorithmischer Entscheidungen im Wertpapierhandelsgesetz (WpHG)

Jenseits der in der vorliegenden Studie betrachteten Anwendungsgebiete spielen algorithmische Entscheidungen bereits heute im Wertpapierhandel eine wesentliche Rolle; zum Teil werden diese binnen Sekundenbruchteilen getroffen. Der Gesetzgeber hat auf diese Entwicklung mit dem Hochfrequenzhandelsgesetz reagiert. Mit diesem Artikelgesetz wurden insbesondere das Börsengesetz, das Wertpapierhandelsgesetz und das Kreditwesengesetz angepasst.

Die neuen Regelungen betreffen die Zuverlässigkeit und Sicherheit der verwendeten Systeme und sollen Störungen des Marktes sowie Marktmanipulationen verhindern.[169] Spezielle Regelungen für Algorithmen, die maschinelle Lernverfahren einsetzen, gibt es nicht. Die Zielrichtung der genannten gesetzlichen Regelungen steht also nicht im Zusammenhang mit dem vorliegend zu untersuchenden Problembereich. Erwähnenswert ist allerdings die Rolle der Aufsichtsbehörde: Gemäß § 6 Abs. 4 WpHG kann die Bundesanstalt für Finanzdienstleistungsaufsicht Informationen über den algorithmischen Handel und die dafür eingesetzten Systeme von Wertpapierdienstleistungsunternehmen anfordern. Organisatorische (Sicherheits-)Maßnahmen sind in § 80 Abs. 2 und

[168] Hacker 2018, Teaching Fairness to Artificial Intelligence, Working Paper, April 2018, S. 25, 28 f., 32.
[169] Dazu ausführlicher der Regierungsentwurf, BT-Drs. 17/11631, S. 1-2.

Dokumentationspflichten in § 80 Abs. 3 WpHG geregelt. Unser Interview mit der Wertpapiergruppe der Bundesanstalt ergab jedoch, dass diese die Algorithmen selbst regelmäßig nicht überprüft.

Das Modell, algorithmische Entscheidungen in besonders risikobehafteten Bereichen von bestimmten technischen und/oder organisatorischen Maßnahmen abhängig zu machen, die von einer Aufsichtsbehörde überprüft werden können, lässt sich grundsätzlich auch auf die vorliegend betrachteten Anwendungsfälle übertragen. Die Erfahrungen, die im Anwendungsbereich des WpHG gemacht wurden, lassen sich aber nicht verallgemeinern. Zum einen ist die durch das WpHG adressierte Gefahrenlage dafür zu spezifisch; zum anderen betreffen die Regelungen des WpHG zum algorithmischen Handel einen ohnehin bereits stark regulierten Sektor, was bei anderen Anwendungen algorithmischer Entscheidungen nicht der Fall ist.

Eine tiefergehende Analyse der Regelungen des WpHG wird daher an dieser Stelle nicht durchgeführt.

6 Regulierung und Standardisierung im internationalen Vergleich

6.1 Übersicht und Kontextualisierung

Das Bewusstsein der Problematik ist in allen analysierten Ländern hoch, sowohl bei politischen Entscheidungsträgern als auch in der Bevölkerung. Dies ist getrieben zum einen durch eine Reihe von medienwirksamen Vorfällen, die die Gefahren und den Missbrauch von Algorithmen auch einer breiteren Öffentlichkeit vermittelten,[170] zum anderen durch die Veröffentlichung einer Reihe populärwissenschaftlicher Studien, die über problematische Einzelfälle hinaus das Gefährdungspotenzial der algorithmischen Gesellschaft hervorgehoben haben.[171] Beispiele von Diskriminierung aufgrund von Geschlecht und ethnischer Herkunft dominieren dabei die Debatte, die so insbesondere in den USA politisch stark polarisierte Positionen annehmen kann. Dies erzeugt zum einen politischen Handlungsdruck, zum anderen aber auch eine Dynamik, die die genaue Ausgestaltung neuer Regulierung erschwert.

In einem Versuch, die dynamische und noch sehr heterogene Diskussion zu systematisieren, strukturieren wir dieses Kapitel nicht nach Rechtsordnungen, sondern stellen exemplarisch Vorschläge vor – einige wie das französische Gesetz zur Fairnesspflicht der Onlineplattformen sind nur in einer Rechtsordnung zu finden, andere typischerweise in der Diskussion in mehreren Ländern. Es ist möglich, die Ergebnisse grob in drei Typen zu kategorisieren: So finden wir Vorschläge, die in erster Linie die Transparenz erhöhen wollen und sich dann typischerweise auf die Regulierung durch den Markt und den freien Wettbewerb verlassen. Ein sehr anderer Ansatz verbietet kategorisch (Formen der) Diskriminierung. Hier kann der Schwerpunkt entweder auf der Analyse der Ergebnisse (a posteriori) sein, mit Sanktionen, wenn sich diese als diskriminierend erweisen, oder auf dem „Input", d.h. einem A-priori-Verbot bestimmter Techniken, oder einem Gebot, nur bestimmte geprüfte Methoden zu verwenden. Parallel zu dieser Unterscheidung finden sich Vorschläge, die sich primär auf den Input in den Algorithmus konzentrieren und etwa durch Datenschutzregeln gewisse Daten gar nicht erst verfügbar machen, während andere sich primär auf das Ergebnis der Analyse und die Verwendung des Ergebnisses konzentrieren. Häufig werden alle drei Ansätze verbunden, insbesondere in Vorschlägen, die nach einer neuen Aufsichtsbehörde verlangen.

[170] Mit weiteren Beispielen: O'Donnell, F.: What we talk about when we talk about fair AI, BBC News Labs, 11.12.2017 [https://bit.ly/2AAkk6t]; Guynn, J. (1. Juli 2015). „Google Photos labeled black people 'gorillas'". USA TODAY; Day, M. (31. August 2016). „How LinkedIn's search engine may reflect a gender bias". *The Seattle Times*; Angwin, Julia et al. „Machine bias: There's software used across the country to predict future criminals and it's biased against blacks." *ProPublica, May* 23 (2016); Derek Hawkins, „Researchers use facial recognition tools to predict sexual orientation. LGBT groups aren't happy", *The Washington Post;* Asian People Are Not Impressed With Their Matches On Google's Museum Selfie Feature, Buzzfeed 17.01.2018; Hayasaki: Is AI sexist? Foreign Policy 16.01.2017; Ian Sample: Computer says no: why making AIs fair, accountable and transparent is crucial, The Guardian 4.11.2017.

[171] Siehe als besonders einflussreich, gemessen am Zitationsindex, Besprechungen und Suchmachinenranking, insbesondere Pasquale 2015, O'Neil 2016, Finn 2017, Finlay 2014, Siegel 2013, Noble 2018, Wachter-Boettcher 2017, Steiner 2013, Dormehl 2014.

Was in der gegenwärtigen Diskussion zur Regulierungsthematik oft vergessen wird, ist wie alt das Problem ist – so gab es bekannte Beispiele für rechtswidrig diskriminierende lernende Algorithmen bereits in den frühen 1980er Jahren. Der Algorithmus, der von 1982 bis 1985 die Zulassung zur St. George's Hospital Medical School kontrollierte, benachteiligte so systematisch Bewerber mit „ausländisch klingenden" Namen aufgrund historischer Entscheidungsmuster.[172] Älter ist sogar die Frage, ob Datenschutzrecht geeignet ist, den Missbrauch von Scoring zu verhindern. Wir diskutierten oben Gesetze aus der Zeit der Bürgerrechtsbewegung in den USA, die diskriminierende Entscheidungen zu verhindern suchen. Gleichzeitig mit dem Erlassen dieser Gesetze wurde der Fair Credit Reporting Act, 15 U.S.C. § 1681, im Jahr 1970 erlassen und 1996, 2003 und 2010 erheblich geändert. Der FCRA begrenzt nicht, welche Informationen von Kreditauskunfteien gesammelt werden, sondern konzentriert sich auf die Beschränkung des Zugangs Dritter zu Kreditdaten für zulässige Gewährleistung der Richtigkeit dieser Daten, Benachrichtigung der Verbraucher über nachteilige Maßnahmen auf der Grundlage solcher Daten und Gewährleistung des Zugangs der Verbraucher zu Daten, und die Möglichkeit, Daten über sich selbst zu korrigieren. Die Anhörungen der Kreditinstitute und des Equifax-Vorläufers Retail Credit Company durch den US-Kongress brachten auch damals schon die gleichen Fragen zum (Miss)brauch von Big Data zur Profilierung, wie wir sie heute finden.[173]

Die Debatte zur rechtlichen Regulierung von „klassischen", d.h. nichtlernenden, Algorithmen durch die öffentliche Hand ist nur wenig älter, mit einem der ersten Beispiele der *Social Security Act 1998* in England, der vollautomatische Entscheidungen über Sozialhilfeansprüche auf eine rechtliche Grundlage stellte und auch in der parlamentarischen Debatte große Bedenken gegenüber vollautomatischen Entscheidungen hervorrief.[174] In der Diskussion zu diesen regelbasierten Algorithmen (Expertensysteme) waren indes die Sorge nicht so sehr diskriminierende, intransparente oder ungerechte Entscheidungen (regelbasierte Systeme sind zumindest in der Theorie hier weniger anfällig), sondern die Frage, ob ein solches Modell der Interaktion zwischen Staat und Bürgern mit Grundprinzipien der Menschenwürde zu vereinbaren ist.[175] Diese Diskussion ist daher nur bedingt auf die gegenwärtige Debatte übertragbar.

Von akademischer Seite gab es seit den 1990er Jahren Warnungen – in einer frühen Studie von Friedman und Nissenbaum identifizierten die Autoren diskriminierende Algorithmen im Gesundheitswesen (Diskriminierung von verheirateten Patienten in der Zuweisung von Pflegeplätzen) und wettbewerbsrechtlich problematische Empfehlungen von Flügen.[176] Diskriminierende Rankings von Suchmaschinen dominierten diese frühen Debatten.[177]

Trotz des schon früh existierenden Bewusstseins über mögliche ethische und rechtliche Probleme führten diese Debatten nicht zu einer effizienteren Regulierung. So ist es möglich, dass mehr als 30 Jahre, nachdem der St. George's Algorithmus aufgrund von Namen Minderheiten diskriminierte, gleiche Vorwürfe gegen Versicherer in Großbritannien von der

[172] Lowry 1988.
[173] Siehe etwa Retail Credit Co. of Atlanta, Ga 1968; für eine historische Diskussion siehe Ohlhausen/Okuliar 2015.
[174] Le Sueur 2016.
[175] So etwa Oliver Letwin MP in der parlamentarischen Anhörung, zu finden als *Standing Committee B, 28 October 1997 (Morning), Keith Bradley MP, Parliamentary Under-Secretary of State for Social Security*.
[176] Friedman/Nissenbaum 1996, S. 48-51.
[177] Siehe z.B. Mowshowitz/Kawaguch 2002, S. 56-60; Introna/Nissenbaum 1999.

Presse gemacht werden konnten.[178] Die Journalisten hatten unter Benutzung verschiedener Namen, aber ansonsten identischer Angaben Quotierungen für eine Fahrzeugversicherung beantragt. „Mohameds" wurden dabei systematisch höhere Raten angeboten als „Johns". Es muss angemerkt werden, dass beide genannten Versicherungen die Vorwürfe strikt zurückweisen. Sie argumentieren, dass die unterschiedlichen Angebote entweder „unvergleichbar" gewesen seien, da sie neben dem Preis auch in den Leistungen unterschiedlich waren, dass die Journalisten die Preisstruktur nicht richtig verstanden hätten oder dass die Verwendung identischer Daten trotz verschiedener Namen den Betrugswarnungsalgorithmus ausgelöst hätte und dies zu unterschiedlichen Resultaten führte.[179]

Wenngleich die Probleme und Diskussionen daher nicht neu sind, und in dieser Hinsicht bislang Regulierung oder Selbstregierung scheinbar wenig Erfolg hatten, zeigt eine genauere Analyse dieses und ähnlicher Fälle aus den USA und Großbritannien auch, wie sich seit 1982 das gesellschaftliche und politische Umfeld geändert hat. Waren es in den 80ern ausschließlich Wissenschaftler mit dem notwendigen Fachwissen, so waren es in diesem und einer Reihe ähnlicher Fälle Enthüllungsjournalisten, die den Algorithmus testeten. Ihre Methoden waren einerseits sehr einfach, schnell und billig, andererseits aber auch nicht sehr systematisch: Einige Faktoren wurden kontrolliert, andere nicht, ohne dass eine besondere Systematik erkennbar wäre. So wurden verschiedene Computer mit verschiedenen IP-Adressen verwendet, um auszuschließen, dass der angebotene Preis das Ergebnis von „personeller Preiskalkulation" war, die außer den groben Kategorien aus dem Antragsformular auch noch Daten der individuellen Onlineinteraktion, wie z.B die Geolokalisierung des Kunden, hinzunimmt.[180] Dies zeigt einerseits, dass die Sorge, dass sich die algorithmische Blackbox prinzipiell der gesellschaftlichen Aufsicht entzieht, unbegründet ist. Insbesondere auch dann, wenn wie im Fall der COMPAS-Enthüllungen[181] zum rassistischen Strafbemessungsalgorithmus Wissenschaftler und Journalisten zusammenarbeiteten,[182] waren die Ergebnisse überzeugend und erzeugten nun auch in der breiteren Öffentlichkeit den Ruf nach staatlicher Regulierung – wie unter anderem die Einreichungen zu den öffentlichen Konsultationen zeigen.

Andererseits offenbaren diese Beispiele auch die Grenzen solcher Analysen durch Privatpersonen: Sowohl in Großbritannien als auch den USA haben die Firmen, die die Algorithmen entwickelten (USA) oder verwendeten (UK), die Schlussfolgerung, dass eine ungerechtfertigte Diskriminierung vorliegt, zurückgewiesen. Dabei traten sie keinen Gegenbeweis an, sondern beschränkten sich darauf, mögliche Fehlerquellen in den Studien aufzuzeigen. Für uns belegt dies die Notwendigkeit standardisierter und extern validierter Prüfungsschemata und Methoden als Minimum.

[178] Ben Leo, Motorists fork out £1,000 more to insure their cars if their name is Mohammed, The Sun, 22. Januar 2018 [https://bit.ly/2Bl2MGE], siehe auch DecisionMarketing, Admiral Insurance hit by big data discrimination claims, 24. Januar 2018 [https://bit.ly/2CBk1tt].

[179] Vic Motune, Are black drivers paying more car insurance?, The Voice vom 17. Februar 2018 [https://bit.ly/2wWlqFM].

[180] Personalisierte Preissetzung ist möglich und wird auch von einigen Unternehmen verwendet, doch zeigt eine Studie für die US-Regierung, dass bislang (Stand 2015) dies noch die Ausnahme zu sein scheint. Siehe Executive Office of the President of the United States (Council of Economic Advisors) 2015: Big Data and Differential Pricing.

[181] Angwin et al. 2016.

[182] Die zentrale Rolle von Journalisten in der Kontrolle von Algorithmen und die Entwicklung neuer Trainingsprogramme und Ressourcen werden diskutiert von Catalina Albeanu: What journalists can do to hold algorithms to account, Journalism UK 14.4.2018 [https://bit.ly/2qsZhL0].

Zu überlegen ist auch, wie das „Nichtbestehen" eines solchen Tests rechtlich zu behandeln ist, und wie wir sehen werden, haben andere Länder unterschiedliche Vorschläge entwickelt – so könnte es wie im oben angeführten Beispiel des *Fair Lending Acts* eine widerlegbare Vermutung der Diskriminierung erzeugen, die dann entweder dadurch entkräftet werden kann, dass Fehler in der Anwendung der Methode gerügt werden, oder sie können den Benutzer verpflichten einen positiven Gegenbeweis anzutreten und die Nichtdiskriminierung des Algorithmus darzulegen (was dann wiederum verschiedene Beibringungs- und Offenlegungspflichten begründen kann). Alternativ kann das Nichtbestehen der Prüfung das Versagen eines Gütezeichens bedeuten, und damit gegebenenfalls das Verbot, in bestimmten Bereichen tätig zu sein.[183]

In der rechtlichen Reaktion auf den angeblich diskriminierenden Versicherungsalgorithmus im UK finden wir beide Modelle – so ist das Unternehmen gegenwärtig Gegenstand einer Untersuchung durch die Financial Conduct Authority, die noch 2016 nach der Analyse eines ersten „Calls for Inputs" abgelehnt hatte, eine Untersuchung zu „Big Data und Versicherungswirtschaft" zu initiieren, da ihrer Ansicht nach die Auswirkung von Big Data Analytics und automatisiertem Scoring für Verbraucher „generell positiv" sei.[184] Hier können die Sanktionen von einem Verbot des spezifischen Algorithmus zu Strafgebühren oder sogar zum Lizenzentzug führen. Gleichzeitig hat der ehemalige Vorsitzende der Equality and Human Rights Commission (EHRC) angedeutet, dass eine Ermittlung der betroffenen Versicherungsunternehmen unter Art. 20 des Equality Acts 2006 durch die EHRC wahrscheinlich sei. Art. 20 verlangt nur einen Anfangsverdacht („suspicion", schwächer als die „reasonable suspicion", die vor der Gesetzesreform 2006 notwendig war und zu sehr vorsichtiger Handhabung führte). Eine Bedingung, die durch Studien wie die durch die BBC und Sun Newspaper durchgeführten problemlos erfüllen. Typischerweise wird die EHRC versuchen, eine Einigung mit dem Unternehmen zu erreichen, bevor eine offizielle Untersuchung durchgeführt wird. In beiden Fällen ist das Ergebnis in der Regel ein verbindlicher Handlungsplan, der zur Beseitigung der Diskriminierung führen soll (Art. 22 und 23 des EA 2006). Möglich ist auch das Erlangen eines Gerichtsbeschlusses zum sofortigen Verhindern diskriminierenden Verhaltens (Art. 24 EA 2006); dies ist auch das Verfahren, wenn ein Unternehmen einen Handlungsplan nicht umsetzt.

Möglich wäre auch eine Ermittlung durch das Information Commissioner's Office, doch gibt es bislang keine Stellungnahme des ICO zu diesem bestimmten Fall. Dies ist für Großbritannien nicht untypisch – einerseits ist das ICO bereit, datenschutzrelevante Verletzungen durch Algorithmen zu untersuchen und zu unterbinden, doch wenn das Problem wie hier primär eine Verletzung von Gleichheitsbestimmungen ist, besteht die Tendenz dies zumindest im ersten Zugriff dem Equality and Human Rights Commissioner oder wenn anwendbar Fachaufsichtsbehörden wie der Financial Service Authority zu überlassen. Im Falle der Versicherungen waren das Problem nicht inakkurate Daten oder eine fehlende Rechtsgrundlage der Datenverarbeitung, und das ICO scheint zurückhaltend zu sein einen Begriff der Fairness und Transparenz zu entwickeln, der zu weit von einem Individualrechtsverständnis abweicht und negative Auswirkungen auf ganze Gruppen hat.

Was die britische Fallstudie auch zeigt, ist, dass diskriminierende Algorithmen häufig in den Zuständigkeitsbereich mehrerer Aufsichtsbehörden und Rechtsgebiete fallen, in diesem Fall ICO, EHRC und FCA. Von diesen hat im Moment die Financial Service Authority das größte

[183] So der Vorschlag eines Algorithm Safety Impact Assessments, vgl. Shneiderman 2016, S. 13538-13540.
[184] Financial Conduct Authority 2016; siehe auch DecisionMarketing, Insurers off the hook as FCA rules out big data probe, 22. September 2016 [https://bit.ly/2O3Fyw6].

Fachwissen, wenn es um technische Fragen geht – hier z.B. ob die Diskriminierung durch Geburtsnamen, die stark mit in der Entscheidungsfindung verbotenen ethnischen Merkmalen korreliert, trotzdem empirisch rechtfertigbar ist.[185] Eine mögliche Algorithmenaufsichtsbehörde muss vermeiden, Aufgaben, die diese Behörden auch weiterhin durchführen müssen, zu duplizieren. Zum anderen bedeutet dies, dass eine einheitliche Methode oder ein universales „Fairness Certificate" wahrscheinlich nicht zu erreichen ist. So sind wie gesehen die Bedürfnisse des EHRC (nur geringer Anfangsverdacht notwendig, nur Diskriminierung gegen sieben im Equality Act genannte Gruppen relevant) anders als die der Financial Conduct Authority. Diese verlangt eine höhere Beweisschwelle, hat aber ein weiteres Aufgabenfeld, das auch die Auswirkung auf den Wettbewerb und die Frage, ob einige Gruppen ganz von erschwinglichen Produkten ausgeschlossen sind, abdeckt. So hat die Financial Conduct Authority in ihrer Analyse des „Calls for Input" zwar geschlossen, das alle Preisermittlungen, die weder risiko- noch nachfragebasiert sind, ein potenzielles Problem darstellen. Eine etwas detailliertere Diskussion des FCA-Berichts findet sich unten.

b) In Hinsicht auf diskriminierende Algorithmen bezieht sich die Besorgnis mit wenigen Ausnahmen auf die Durchsetzung und Durchsetzbarkeit bestehender Diskriminierungsverbote. Nur vereinzelt finden sich Stimmen, die argumentieren, dass algorithmische Diskriminierung ein neuer Diskriminierungstyp ist, der der gesonderten Regelung bedarf. In den USA sind solche Diskussionen zusätzlich durch die extensive Interpretation der verfassungsrechtlich garantierten Redefreiheit eingeschränkt, die über die Jahre mehr und mehr auch auf kommerzielle Rede ausgedehnt wurde. Die genaue Einordnung von algorithmischen Entscheidungen als „Rede" bleibt dabei umstritten[186], kann aber den gesetzgeberischen Handlungsspielraum auch im Verbraucherschutz weitgehender einschränken, als dies in Deutschland der Fall wäre.[187]

c) Wir fanden nur wenige Stimmen, die fordern, für algorithmische Entscheidungen den Haftungstyp zu ändern und insbesondere verstärkt verschuldungsunabhängige Haftung zu verwenden, wenn dies nicht auch schon der Haftungstyp für die manuelle Entscheidung ist – in dieser Hinsicht unterscheidet sich die Debatte zu algorithmischen Entscheidungen in den Problemszenarien von der parallel stattfindenden Diskussion über die Haftung von physisch implementierten Algorithmen (Robotik), in der derartige Ideen weiter verbreitet sind.[188] Die Entscheidung Algorithmen zu verwenden wird als eine zurechenbare Entscheidung ihrer Besitzer begriffen, die diese dadurch weder besser- noch schlechterstellen soll – in Fortführung eines alten Gedanken aus der Internetregulierung, was menschlichen Entscheidungsfindern erlaubt/verboten ist, soll dies auch auf Algorithmen anwendbar sein. Was Algorithmen zu einem Problem der Regulierung macht, ist damit primär die

[185] So hatten Versicherer in einem früheren Fall im Jahr 2010, der damals auch von der Verbraucherschutzorganisation *Which* aufgedeckt wurde, argumentiert, dass Diskriminierungen gegen im Ausland aufgewachsene Fahrer (unabhängig von ethnischer Herkunft oder Nationalität) belegbar ein größeres Risiko darstellen [https://bit.ly/2N3PngS].
[186] Siehe z.B. Benjamin 2012, S. 1445; Massaro/Norton 2015, S. 1169.
[187] Zum potenziellen Konflikt zwischen „First Amendment"-Schutz für algorithmische Rede und Verbraucherschutz siehe Balkin 2017.
[188] So etwa die EU Motion For A European Parliament Resolution with Recommendations to the Commission on Civil Law Rules on Robotics (2015/2103(INL); mit vergleichender Analyse europäischer Ansätze: Final Report Summary - ROBOLAW (Regulating Emerging Robotic Technologies in Europe: Robotics facing Law and Ethics), rechtsvergleichend siehe auch Kelley et al. 2010, S. 1861-1871.

Durchsetzung bestehender Normen, und der Beweis der Normverletzung, nicht das materielle Recht.

d) Bislang allerdings hat diese Besorgnis noch nicht zu abgeschlossenen Reformprojekten außerhalb des Datenschutzrechts geführt. So haben wir keine Beispiele gefunden, in denen es zu einer Änderung des materiellen Verbraucherschutzrechts durchgeführt. Frankreich, als eine mögliche Ausnahme, hat zwar einen generellen „Fairnessparagrafen" in Sec 3 Art. 13 des neuen *Loi n°2016-1321 pour une République numérique* (Digitale-Republik-Gesetz) eingeführt, der dem Verbraucherschutz dienen soll. Wie wir sehen werden, behandelt dieser aber nicht die Szenarien, die im Zentrum dieser Studie stehen, sondern reguliert primär Rankingalgorithmen auf Onlineplattformen und Onlinemärkten.

Vergleichsweise weiter vorangeschritten sind Vorhaben, eine „Algorithmenbehörde" zu errichten. Dies ist etwa im Haushalt der britischen Regierung vorgesehen, der ein *„Centre for Data Ethics and Innovation"* vorschlägt, dessen genaue Aufgaben und Befugnisse am 17.1.2018 Gegenstand einer parlamentarischen Debatte waren.

Auch diese sind aber noch nicht umgesetzt worden, so dass Erfahrungen mit ihrer Anwendung, ihrer Effizienz, ihren Kosten und ihrer öffentlichen Akzeptanz weiterhin fehlen. Indirekt gelernt werden kann aber durch die Anhörungen und Konsultationen, die auf breiter Basis stattfanden und auch Bedenken aus Industrie und rechtlicher Praxis aufgegriffen haben.

e) Auch wenn es weitgehend Übereinstimmung darüber gibt, wie problematische algorithmische Entscheidungen lebensweltlich aussehen (die Arten von Szenarios, die diskutiert werden, sind denen in unserer Studie sehr ähnlich), gibt es keinen Konsens darüber, welche Rechtsgebiete primär betroffen sind. Dies führt auch dazu, dass es keinen Konsens gibt, ob es rechtlicher Änderungen bedarf, einer besseren Selbstregulierung der Industrie (unter Umständen durch flankierende Maßnahmen, um einen effektiven Markt sicherzustellen), technologischer Lösungen (gegebenenfalls mit flankierenden Maßnahmen zur Standardisierung durch ISOs oder ähnliche Kitemarks), oder ob prozessrechtliche Schritte (einschließlich Kostenentscheidungen) einen besserer Weg darstellen. In den Antworten zu diesen Fragen zeigen sich die Auswirkungen von historisch gewachsenen und gesellschaftlichen Unterschieden zwischen den Rechtsordnungen besonders deutlich.

6.2 Klassifizierung der Debatten

Wie oben unter e) gesehen, gibt es bislang noch keinen Konsens darüber, wie das Problem der algorithmischen Diskriminierung begrifflich einzuordnen ist. In einer ersten Annäherung können die Vorschläge in Gruppen zusammengefasst werden.

6.2.1 Analoge Anwendung bestehenden Rechts

Es sollte nicht als gegeben betrachtet werden, dass die neue Technologie rechtlichen Handlungsbedarf erzeugt. Gerade in Fragen der Technikregulierung lenkt häufig die oberflächliche Neuheit der Technologie davon ab, die zugrundeliegenden und häufig sehr viel älteren gesellschaftlichen Probleme und Fragestellungen wiederzuerkennen. Wir sehen dies in einem eng verwandten Fragegebiet, der Regulierung der Robotik als verkörperter KI. Hier fluktuiert die Diskussion zwischen Positionen, die eine radikal neue Problemlage

identifizieren und ihr mit neuen rechtlichen Konzepten wie der E-Person begegnen wollen,[189] und solchen, die vor allem die Ähnlichkeit mit altbekannten Problemen wie etwa der Verantwortlichkeit für Hunde und Pferde hervorheben[190] und auf eine analoge Anwendung von Begrifflichkeiten aus dem 19. Jahrhundert – oder sogar der römischen Republik[191] – setzen.

In dem Ausmaß, in dem verschiedene Rechtsordnungen eine „mentalité" oder kognitive Grundeinstellung haben, tendieren Common-Law-Rechtsordnungen zu letzterem Ansatz, mit einem (im Abnehmen befindenden) Misstrauen gegen gesetzliche Lösungen und stärkerem Vertrauen in die Fähigkeit der Gerichte, im Wege der Analogie angemessene Lösungen im Einzelfall zu finden. Auch haben die USA und das UK wohl eine größere Bereitschaft, marktbasierte Ansätze und Selbstregulierung in den Vordergrund zu stellen. Man sollte diese Unterschiede aber nicht überbetonen, und wie oben ausgeführt sind auch in diesen Ländern Diskriminierungsverbote durch formale Gesetze eingeführt worden.

Gegeben diese „philosophischen" Vorentscheidungen, kann man trotzdem aus den Erfahrungen dieser Rechtsordnungen Argumente für oder gegen einen Handlungsbedarf finden? Dies führt zu einem Paradox: Das in der Diskussion prominenteste Problem mit lernenden Algorithmen ist ihre fehlende Transparenz. Wenn es nun nur wenige identifizierte Fälle von rechtswidrigem Entscheiden gibt, heißt dies, dass bestehende Gesetze dem Missbrauch effizient vorbeugen oder dass sie so ineffizient sind, dass weitgehender Missbrauch gar nicht mehr entdeckt werden kann? Die öffentliche Diskussion ist in starkem Maße durch die Identifizierung von missbräuchlichen und diskriminierenden Algorithmen geprägt worden. Heißt dies, dass anders als befürchtet Ex-post-Analysen von Algorithmen leicht möglich und effizient sind, oder sind diese Fälle nur nichtrepräsentative Beispiele eines viel fundamentaleren Problems?

Trotz dieser methodischen Probleme gibt uns eine rechtsvergleichende Analyse zumindest einige neue Datenpunkte. Zum einen ist wie oben angeführt das Problembewusstsein über Rechtsordnungen und rechtliche Traditionen hinweg sehr hoch, und wie wir unten sehen werden, ist es sehr wahrscheinlich, dass die USA und das UK schon sehr bald durch Gesetzgebung neue Institutionen zur Regulierung von Algorithmen einführen werden.[192] Wenn sogar in Rechtsordnungen, die tendenziell unternehmerfreundlich und gesetzeskritisch sind, starker Handlungsbedarf besteht, sollte dies *a fortiori* auch für kontinentaleuropäische Rechtsordnungen gelten.

Andererseits gibt es durchaus positive Erfahrungen mit der rechtlichen Behandlung von Maschinellem Lernen. Oben besprachen wir den US *Fair Lending Act*, der Diskriminierungsschutz vorsieht (laut einer Industriequelle hat sich die Zahl der verwendeten Kriterien seit 1990 von 15 auf 100 erhöht) – doch scheint dies nicht dazu geführt zu haben, dass bestimmte Gruppen ganz vom Zugriff auf Kreditentscheidungen durch einen Auditmechanismus ausgeschlossen werden – und zumindest nicht offensichtlich schlechter

[189] Szollosy 2017, S. 1-6.
[190] Schaerer 2009, S. 72-77.
[191] Pagallo 2010, S. 397-404.
[192] Für das UK mit weiterer Diskussion unten: Committee sets the agenda for new algorithmic ethics agency, www.parliament.uk, 23. Mai 2018 [https://bit.ly/2s3D73u]; für die USA siehe etwa diesen Gesetzesvorschlag: United States Cong. Senate. FUTURE of Artificial Intelligence Act of 2017. 115th Cong. 1st sess. S.2217 [https://bit.ly/2zcUQJZ]. Siehe auch Lawmakers introduce bipartisan AI legislation, The Hill, 12. Dezember 2017 [https://bit.ly/2AOJJpx].

mit algorithmischen Entscheidungen zurecht kommt als mit den menschlichen oder mechanisch-regelbasierten Ansätzen.

In Großbritannien hat die oben erwähnte Studie der *Financial Conduct Authority* zu Big Data im Versicherungsgewerbe zu einem ähnlichen Ergebnis geführt, obgleich die empirische Grundlage eingeschränkt war (27 Antworten zum „*Call for Evidence*“ durch Industrie und Konsumentengruppen, eine Reihe von Round Tables und die Analyse der Daten von 2 Preisvergleichsseiten). Zum einen können Big Data und Maschinelles Lernen alte Probleme verstärken und beschleunigen, auch wenn die Analyse keine fundamental neuen Problemszenarien identifizierte. Zum anderen aber kann sie auch den Wettbewerb erhöhen und gerade Konsumenten mit untypischen Risikoprofilen angemessenere und niedrigere Quotierungen geben.

Marktsegmentierung könnte, so die FCA, im schlimmsten Fall dazu führen, dass Bevölkerungsgruppen von notwendigen (und erschwinglichen) Versicherungen ganz ausgeschlossen werden. Hier sieht die FCA aber vor allem die Regierung in der Pflicht – nicht Big Data und Algorithmen sind das Problem, sondern gesamtgesellschaftliche Ziele, die am besten in Zusammenarbeit zwischen FCA, Industrie und Regierung erreicht werden; hier weist die FCA insbesondere auf Flutversicherungen und Versicherungen für Behinderte hin. Sie könnte auch dazu führen, dass der „Solidaritätsgedanke“ und das Pooling von Risiko unterminiert werden. Die FCA identifiziert insbesondere Preisentscheidungen, die weder durch eine Risikoevaluierung noch durch Kosten motiviert sind – Beispiele sind etwa Browserverhalten, das anzeigt, dass der Kunde keine Preisvergleiche macht, oder Profile, die andeuten, dass sich der Kunde wahrscheinlich beschweren wird. Zusammenfassend findet die Studie, dass zwar einerseits Profile genauer werden oder genauere Risikoprofile generell zu höheren Kosten für Versicherte geführt hätten. Im Gegenteil, so die FCA, deuten die Daten darauf hin, dass es zwar wie immer Gewinner und Verlierer gibt, generell aber die Auswirkungen von Big Data und Machine Learning für Verbraucher positiv waren und gerade auch durch Preisvergleichsseiten zu erhöhtem Wettbewerb führten.[193]

Die FCA weist in diesem Zusammenhang auch auf die Erfahrung in den USA hin. In allen Staaten (mit Ausnahme von Illinois) gelten Gesetze, nach denen „die Prämien nicht unangemessen, übermäßig oder unfair diskriminierend sein dürfen“. Jeder Staat hat bei der Umsetzung des Gesetzes zur Preisgestaltung einen etwas anderen Ansatz gewählt. Die Mehrheit der Bundesstaaten hat bestimmte Formen der Preisoptimierung durch „Bulletins“, eine Form von Richtlinien, definiert und in der definierten Form generell verboten. So gab etwa im Februar 2015 das Ohio DOI das Bulletin 2015-01 heraus, dass Preisoptimierung als „die Erfassung und Analyse von Daten in Bezug auf zahlreiche Charakteristiken für einen bestimmten Versicherungsnehmer“ definiert, „die nicht mit dem Verlustrisiko oder Kosten in Zusammenhang stehen". Als Beispiel werden die Preiselastizität der Nachfrage angegeben, oder Verhaltensprofile, die vorhersagen, wie viel von einer Preiserhöhung ein bestimmter Versicherungsnehmer tolerieren würde, bevor er den Versicherer wechselt. Maryland, Ohio, Florida, Kalifornien und Vermont folgten diesem Modell. Andere Staaten sind zurückhaltender, so hat das New York Department of Financial Services (NYDFS) in einem

[193] Dies ist ein Argument, das gerade auch in den USA Unterstützung gefunden hat: In diesem Ansatz sollten wir Verbraucher nicht nur als passive Opfer von Algorithmen sehen, sondern stattdessen den „algorithmischen Verbraucher“ stärken, der diese Technologien auch zu seinem Vorteil nutzen kann. Siehe z.B. Gal/Elkin-Koren 2017, S. 309; sehr viel skeptischer Stucke/Ezrachi 2017. Digitale Assistenten als Gehilfen von Verbrauchern konnten im Rahmen dieser Studie nicht diskutiert werden, doch werfen auch sie verbraucherschutzrechtliche Fragen auf – insbesondere über ihre „Loyalität“ zu ihrem Benutzer.

Brief an die Versicherer vom 18.3.2016 angedeutet, dass einige Formen der Preisoptimierung Gesetze gegen unfaires Verhalten verletzen könnten, aber den Versicherungsunternehmen bislang nur aufgetragen, weitere Studien und Informationen zu liefern. Einige Staaten halten existierende Regulierungen für ausreichend, sofern die Preisoptimierung eine oder mehrere spezifischer Bedingungen (*constraints*) erfüllt, so zum Beispiel, dass die Abweichung zwischen versicherungsmathematischem Risiko und dem optimierten Preis nur innerhalb bestimmter Bandbreiten variiert, immer nur zu Gunsten des Kunden variiert, und/oder bestimme Offenlegungspflichten erfüllt.[194]

Die Nationale Vereinigung der Versicherungskommissare (NAIC) – eine US-Aufsichtsbehörde, die von Versicherungsaufsichtsbehörden aus allen 50 Staaten gegründet und regiert wird – hat in einer diesbezüglichen Arbeit von 2015 empfohlen, die Berücksichtigung folgender Faktoren zu verbieten: Preiselastizität der Nachfrage, Bereitschaft zum Abschluss einer Versicherungspolice und die Neigung eines Versicherungsnehmers, Fragen zu stellen oder Beschwerden einzureichen. Ähnliche Verbote wurden auch für das UK als wünschenswert vorgeschlagen, wären aber auch dort „agnostisch" bezüglich der verwendeten Entscheidungsmethode und nicht algorithmenspezifisch, auch wenn es die Verwendung von Algorithmen ist, die es sehr viel leichter macht, diese Faktoren zu berücksichtigen. Von diesem Vorschlag abgesehen sieht die FCA aber keinen Handlungsbedarf.

Diskussionen in den USA folgen einer ähnlichen Linie. So steht in einem Bericht des Weißen Hauses: „Viele Unternehmen verwenden bereits Big Data für gezieltes Marketing und einige experimentieren mit personalisierten Preisen. Auch gefördert durch […] Preisdiskriminierung, basierend auf breiten demographischen Kategorien, hin zu personalisierten Preisen." Doch fand diese Studie auch, dass sich die Methoden noch im Experimentierstadium befinden.[195]

Ungeachtet der obigen Beispiele scheint eine personalisierte Preisgestaltung relativ selten zu sein. In einer US-fokussierten Studie von Narayanan (2013) kommt der Autor zu dem Ergebnis: *„The mystery about online price discrimination is why so little of it seems to be happening"*. Ein Grund kann sein, dass Unternehmen feindliche Reaktionen der Öffentlichkeit befürchten. Sowohl das Office of Fair Trading im Jahr 2010[196] als auch das Exekutivbüro des Präsidenten der Vereinigten Staaten im Jahr 2015[197] kommen zu dem Ergebnis, dass transparente und effiziente Märkte ausreichend sind, um problematische Preisdiskriminierung zu verhindern. Von akademischer Seite war dies bereits 2009 von Odlyzko vorhergesagt worden. „The main constraint on price discrimination comes from society's dislike of the practice." Er fügte hinzu: „What forms of price discrimination society will accept. So we should expect experimentation, hidden as much as sellers can manage, but occasionally erupting in protests, and those protests leading to sellers pulling back, at least partially. And occasionally we should expect government action, when the protests grow severe."[198] Die Reaktion auf die Anschuldigungen gegen die britischen Versicherer zeigt, dass diese Einstellung auch 2018 vorherrschend ist – einer der Versicherer erwägt Klage gegen die Journalisten, da der Reputationsschaden signifikant ist. Sowohl die FCA in Großbritannien als auch die OFT in den USA gehen im Moment noch davon aus, dass

[194] Casualty Actuarial and Statistical (C) Task Force, Price Optimization White Paper, 19. November 2015 [https://bit.ly/1NyfROF].
[195] Executive Office des Präsidenten der Vereinigten Staaten 2015, S. 2-4.
[196] Office of Fair Trading 2010.
[197] Executive Office of the President of the United States (Council of Economic Advisors) 2015.
[198] Odlyzko 2009.

staatliches Handeln (*governmental action*) nur vereinzelt und dann typischerweise als „Versicherer des letzten Auswegs“ in Erscheinung tritt.

Schlussendlich sei auf eine Gerichtsentscheidung aus Finnland hingewiesen.[199] Das finnische Nichtdiskriminierungs- und Gleichstellungstribunal hatte einen Fall geprüft, in dem ein Kreditantrag aus statistischen Gründen aufgrund des Geschlechts, der Sprache, des Alters und des Wohngebiets des Antragstellers abgelehnt worden war. Das Gericht brachte vor, dass das System Menschen, die andere Sprachen als Finnisch und Schwedisch sprechen, benachteilige. Das Schiedsgericht befand, das Finanzinstitut habe „diskriminierendes statistisches Profiling“ vorgenommen, und hat das Finanzinstitut angewiesen, statistische Methoden nicht „diskriminierend“ zu benutzen. Die Verwendung Künstlicher Intelligenz bei Kreditentscheidungen wurde nicht verboten, aber eine Geldstrafe von 100.000 Euro auf Bewährung soll sicherstellen, dass es zu keinen weiteren Diskriminierungen kommt. Das Gericht ist eine unabhängige, vom Kabinett ernannte Stelle. Es kontrolliert die Gleichheit sowohl im privaten als auch im öffentlichen Sektor, ausgenommen innerfamiliäre Angelegenheiten und Religion. Es kann Verbote aussprechen, aber keine Entschädigung verlangen. Vor dem Verfahren hatte der finnische Gleichstellungsombudsmann Gespräche mit dem Finanzinstitut geführt, um eine für beide Seiten annehmbare Lösung zu finden.

Dieses Kapitel behandelte Erfahrungen mit der Anwendung bestehender Gesetze auf automatische Entscheidungen. Common-Law-Rechtsordnungen, aber auch ein Gericht aus der nordischen/kontinentaleuropäischen Familie haben gezeigt, dass dies zumindest nicht unmöglich ist. Dies sollte keine Überraschung sein: Das am meisten in der Diskussion genannte Problem mit algorithmischer Entscheidungsfindung ist der Mangel an Transparenz, die Angst vor der Blackbox. Aber auch menschliche Entscheidungen sind zumindest teilweise opak. So verlangen wir etwa von Richtern nur, dass sie objektiv nachvollziehbare Entscheidungsgründe angeben und dass sie offensichtliche diskriminierende Einflüsse erkennen und vermeiden. Wir verlangen aber keine psychologische Evaluierung, die subtile Vorurteile und Vorlieben während der formativen Jahre aufdeckt. Ähnlich können Ansätze zum Algorithmenauditing verstanden werden. Sie haben sich bei der Identifizierung von direkter und häufig auch den offensichtlicheren Beispielen indirekter Diskriminierung durchaus bewährt, auch wenn es schwer sein wird, sehr indirekte Diskriminierung, die Vorhersage von verbotenen Merkmalen aufgrund einer Kette von Abhängigkeiten zu erlaubten Merkmalen gegen sehr spezifische Gruppen mit sich überschneidenden Merkmalen (z.B. eine Kombination aus Alter, Ethnizitäten und Geschlecht), zu identifizieren. Die Frage ist, wie weit diese Analogie geführt werden soll. So scheint es auch im internationalen Vergleich breiten Konsens zu geben, dass die Verwendung von Algorithmen nicht dazu führen darf, bestehende Antidiskriminierungsgesetze zu umgehen. Was Menschen verboten ist, darf Algorithmen nicht erlaubt sein.[200] Die Frage aber bleibt, ob der Umkehrschluss auch gelten soll. Menschen erwerben ihre Vorurteile durch komplexes soziales Lernen über Jahre hinweg, so dass es unmöglich ist, sicher zu sein, ob eine gegebene Entscheidung das Ergebnis einer ungerechtfertigten, erlernten Generalisierung war. Nun „lernen“ zwar auch Algorithmen, doch wie wir im technischen Teil gesehen haben, ist dies im Maschinellen Lernen nicht ganz

[199] Daily Finland, 26.04.2018: Credit decisions - Discrimination through artificial intelligence banned, zuletzt online am 08.08.2018 [https://bit.ly/2NtnkqJ].

[200] Entgegenstehende Meinungen sind selten – so aber argumentiert etwa Gal/Elkin-Koren (2017), dass „algorithmische Verbraucher“ so viel besser Diskriminierung durch andere Algorithmen vermeiden können und rechtliche Diskriminierungsverbote abgeschwächt werden können.

das Gleiche, und Trainingsdaten und Modelle können im Prinzip zugänglich gemacht und analysiert werden.

Damit kommt dieses Kapitel zu einem vorsichtig-optimistischen Ergebnis: Unabhängig von rechtlicher Tradition ist ein Regulierungsbedarf identifiziert worden – Unterschiede bestehen im Ausmaß und in der Methodik. Regulierung ist aber auch *möglich*, die Blackbox nicht Schicksal, wie es einige Anhänger des Technikdeterminismus behaupten. Selbst relativ einfache Methoden der Analyse der Ergebnisse von Algorithmen und etwaiger problematischer Verhaltensmuster und Methoden, die sich in der Analyse nichtalgorithmischer Diskriminierung bewährt haben, erlaubten die Identifizierung diskriminierender Algorithmen und angemessene rechtliche Sanktionen.

6.2.2 Neue gesetzliche Fairnessgebote und eine „Lex algoritmica“

Diametral entgegengesetzt sowohl in ihrer Einschätzung der Problemlage als auch in ihren Lösungsvorschlägen sind Ansätze, die gesetzlich gänzlich neue und technologiespezifische Diskriminierungsverbote oder Fairnessgebote einführen. Vorschläge dieser Art sind primär in der akademischen Diskussion zu finden.[201]

Konkrete Umsetzungen sind selten, ein potenzielles Beispiel findet sich aber im neuen französischen *Digitalgesetz* oder *Gesetz für eine Digitale Republik* (Loi n°2016-1321 pour une République numérique).[202] Abschnitt 3 des Gesetzes befasst sich mit „Plattformfairness.“

Artikel 13: Im Rahmen der jährlichen Erhebung des Conseil d‘Etat zu digitalen Technologien und Grundrechte („Numérique et droits fondamentaux“) im Jahr 2014 wurden Plattformen im Wesentlichen als Inhaltslisten und Ranglisten von Dritten definiert. Dies beinhaltet Suchmaschinen, soziale Netzwerke, Marktplätze usw. Schon die ersten Studien zu algorithmischer Diskriminierung aus den 1990ern hatten, wie wir gesehen haben, Plattformen als aktive Vermittler identifiziert, deren Rolle bei weitem nicht neutral ist. Da einige dieser Plattformen sehr einflussreich geworden sind, kann es durchaus zu Verstößen gegen bestehende Rechtsvorschriften kommen, insbesondere gegen die Fairness gegenüber den Verbrauchern, aber auch gegenüber Unternehmen, deren Ranking nach unten manipuliert wurde.

Artikel 13 sieht vor, dass in Artikel L.111-5-1 des Verbraucherschutzgesetzes eine Definition von Onlineplattformen aufgenommen wird, und verpflichtet Plattformen zur Fairness gegenüber Benutzern. Die Verpflichtung umfasst ihre allgemeinen Nutzungsbedingungen und die Methoden für die Auflistung, das Ranking / die Klassifizierung und das Delisting von Onlineangeboten. Die Offenlegung der „Methoden der Auflistung“ reflektiert dabei das „Recht zur Erklärung der Logik automatischer Entscheidungen“ in Art. 15 der DSGVO, geht aber in mehrfacher Weise darüber hinaus. So schützt er neben natürlichen Personen auch Unternehmen und deckt auch Ranking-Entscheidungen ab, die nicht auf persönlich identifizierbaren Daten beruhen. Wie detailliert diese Offenlegung aber sein muss, und insbesondere ob sie ausreichend sein muss, um durch Dritte nachvollziehbar und überprüfbar zu sein, ist noch unklar.

[201] So etwa für Anwendungen im öffentlichen Sektor Citron 2007.
[202] La République française, Explanatory Memorandum, zuletzt online am 08.08.2018 [https://bit.ly/2x2k774].

Artikel 13 legt außerdem fest, dass die Plattformen alle vertraglichen Beziehungen oder Eigentumsverhältnisse mit den gelisteten Personen klar anführen und benennen müssen, ob diese Personen eine Entschädigung erhalten, und gegebenenfalls die Auswirkungen auf das Ranking von Inhalten und Diensten.

Um ein Ökosystem von Plattformen zu entwickeln, sieht Art. 14 vor, dass eine öffentliche Behörde einen effizienten und fairen Wettbewerb zwischen den Betreibern gewährleisten kann, wobei Innovation und Marktexpansion im Vordergrund zu stehen haben.

Diese Behörde wird beauftragt, Prinzipien über die Informationen zu ermitteln, die die Plattformen den Verbrauchern zur Verfügung stellen müssen, um eine einfache Vergleichbarkeit zu ermöglichen. Die Behörde entscheidet über die Genauigkeit und das Format dieser zu erhebenden Informationen sowie über die Indikatoren für die Bewertung und den Vergleich dieser Praktiken. Schließlich kann sie, wenn sie die bereitgestellten Informationen für nicht ausreichend hält, die Daten sammeln und verbreiten, die erforderlich sind, um die Verbraucher auf dem Laufenden zu halten und ihnen Vergleiche zu ermöglichen. Der Zweck dieses ersten Schritts besteht darin, die Praktiken dieser Plattformen besser zu objektivieren und die Diskussionen, insbesondere auf EU-Ebene, über potenziell restriktivere wirtschaftliche Regelungen zu befördern.

In vielerlei Hinsicht ist dieser Gesetzesvorschlag mit seinem Fokus auf Onlineplattformen und deren Rankings enger als die Thematik dieses Berichts, in anderer Hinsicht geht er darüber hinaus und schützt auch Unternehmen und andere Marktanbieter. Details werden erst zu einem späteren Zeitpunkt durch eine neue Aufsichtsbehörde entwickelt werden, und es ist wahrscheinlich, dass der Begriff der „loyauté“ unter Umständen inadäquat als „Fairness“ übersetzt, auch durch Gerichte interpretiert werden muss. Relevant ist hier insbesondere der Rapport no 3119, der Assemblée Nationale vom Oktober 2015, *Numérique et libertés: un nouvel âge démocratique*[203], der die Motivation und das Ziel der Gesetzgebung kurz beschreibt. „loyauté“ ist dabei allerdings nur negativ beschrieben als Verhalten, das über bloße Neutralität hinausgeht. Ansonsten akzeptiert der Bericht, dass es zwischen dem rechtlichen und informationswissenschaftlichen Verständnis noch Unterschiede gibt, die erst durch zukünftige Forschung und Diskussion geschlossen werden können.

Trotzdem können einige Schlussfolgerungen gezogen werden:

a) Andere Abschnitte des Gesetzes setzten direkter Ideen der DSGVO um, insbesondere Datenübertragbarkeit. Dies zeigt an, dass der französische Gesetzgeber die DSGVO alleine als unzureichend auffasst, um eine diskriminierende Anwendung von Algorithmen zu verhindern. Dies deckt sich mit den Diskussionen in den USA und Großbritannien – Datenschutz ist Teil der Lösung, aber nicht die gesamte Lösung, und zusätzliche Regulierung ist notwendig.

b) Sehr ungewöhnlich war das Gesetzgebungsverfahren, das eine Form des „Crowdsourcing“ oder ein besonders offenes und interaktives Konsultationsverfahren verwendete. Ein Onlineportal erlaubte, Änderungsvorschläge direkt als Text einzugeben, zu rechtfertigen, mit anderen zu diskutieren und über sie abzustimmen. Über 4000 Vorschläge wurden gemacht. Art. 13 und 14 erregten relativ geringe Debatten und die

[203] Französische Nationalversammlung: Commission de réflexion et de propositions sur le droit et les libertés à l’âge du numérique, zuletzt besucht am 26.07.2018 [https://bit.ly/2MgfdsI].

Änderungsvorschläge waren primär auf bessere Lesbarkeit ausgerichtet.[204] Dies zeigt eine generelle Akzeptanz des Grundprinzips in der Bevölkerung und, soweit man aus der Abwesenheit von Gegenvorschlägen Rückschlüsse ziehen kann,[205] auch in der Industrie. Nur ein Industrievertreter kommentierte zu diesen Paragrafen: die *Association des Fournisseurs d'Accès et de Services Internet*, der Verband der Internet Service Provider, der eine explizitere und eingeschränktere Definition des Begriffs „Onlineplattform" vorschlug. Dies deckt sich durchaus mit den (zugegebenermaßen nicht systematisch analysierten) Stellungnahmen der Industrie in Großbritannien und auch (doch hier etwas gemischter) den USA. Auch die Industrie sieht die Vorteile und die Notwendigkeit, das Vertrauen der Verbraucher in algorithmische Entscheidungen zu erhöhen, und akzeptiert, dass dies zumindest teilweise durch gesetzliche Regeln erreicht werden muss.[206] Dies ist sehr stark der Tenor sowohl der französischen Initiative als auch des unten diskutieren britischen Vorschlags einer neuen Behörde. Beide betten die Initiativen direkt und explizit in nationale KI-Strategien ein und betonen die positiven Auswirkungen effizienter Regulierung auf die notwendige öffentliche Akzeptanz – eine Position, die weitgehend von der Industrie geteilt zu werden scheint, wobei allerdings bedacht werden muss, dass es bislang nicht möglich ist, die Kosten dieser Vorschläge für die Wirtschaft zu quantifizieren.

c) In unserer Systematik ist der französische Vorschlag primär an Transparenz interessiert und vertraut daher letztendlich auch auf die regulatorische Wirkung des Marktes. Obgleich es noch unklar ist, wie substantiell die neu zu gründende Behörde oder die Gerichte den Fairnessbegriff interpretieren werden, deuten die offizielle Begründung des Gesetzes und auch der Kontext der anderen Paragrafen darauf hin, dass in erster Linie Offenheit und ein Manipulationsverbot gemeint ist, nicht so sehr materielle Ungerechtigkeit oder Diskriminierung. Dies setzt effiziente Märkte voraus.

d) Obwohl der Vorschlag einer der wenigen ist, die direkt materielles Recht ändern und einen neuen Anspruch erschaffen, ist auch dieses Modell letztlich auf die Einrichtung einer neuen Aufsichtsbehörde ausgerichtet. Dies passt auch zu a) oben: Die Aufgaben sind zu unterschiedlich und der Schutzzweck zu unterschiedlich, als dass dies durch die französischen Datenschutzbehörden übernommen werden könnte. Das bedeutet, das neben der datenschutzrechtlichen Zertifizierung nach Art. 42 und 43 DSGVO ein neues Zertifikat, ein „Zertifiziert Transparent", eingeführt und durch eine neue Behörde vergeben werden muss.

[204] Siehe etwa diesen Änderungsvorschlag im "Projet de loi pour une République numérique" [https://bit.ly/2wZLQp3]; ein Beispiel einer inhaltlichen Änderung, das vorschlägt das „fair und transparent" zu „fair, transparent und korrekt" zu erweitern (5 Zustimmungen, keine Gegenstimme), kann hier gefunden werden: [https://bit.ly/2oUdUqh].

[205] Es ist natürlich denkbar, dass Industrievertreter andere Kommunikationskanäle bevorzugen, insbesondere wenn es darum geht öffentlich und unter eigenem Namen einem Vorschlag zu widersprechen, der sie zu größerer Fairness und Transparenz gegenüber den eigenen Kunden verpflichtet.

[206] Typisch in dieser Hinsicht sind die Beiträge der Industrievertreter beim *Second Meeting of the All-Party Parliamentary Group on Artificial Intelligence* (APPG AI) [https://bit.ly/2CMmmBY]; siehe auch No 6 der Empfehlungen dieser Gruppe, nach Anhörung von 1309 Experten und Stakeholdern [https://bit.ly/2QgplFj]; ähnlich [https://bit.ly/2oWjEQq] mit gegensätzlichen Ansichten zweier Industrieexperten.

6.2.3 Datenschutzbasierte Ansätze

Die Rolle des Datenschutzes als Steuerungsmechanismus ist insbesondere in der europäischen Diskussion prominent vertreten[207] und findet auch zunehmend Interesse in den USA.[208] So hat in Großbritannien das Büro des Datenschutzbeauftragten in einem Bericht zu Big Data und AI einerseits auf die eigenen Erfahrungen und Kompetenzen hingewiesen und ausdrücklich eine Führungsfunktion für die Diskussion verlangt, andererseits aber auch hervorgehoben, dass nicht jedes unerwünschtes Verhalten von Algorithmen automatisch die Verwendung persönlicher Daten beinhaltet.

Das *„Data Ethics Framework"* des Cabinet Office von 2014 bezieht zum ersten Mal die Verwendung von automatischen und semiautomatischen Entscheidungsmethoden als einen Parameter des verpflichtenden *Ethics Risk Assessments* für datengetriebene Innovation und neue datengestützte Anwendungen durch die öffentliche Verwaltung ein; das Ergebnis ist im Wesentlichen ein Privacy Impact Assessment mit Zusatzfragen zur Automatisierung der Entscheidungsfindung.[209]

Auch in den USA ist der Gedanke, Datenschutzrecht als Regulierungsmedium zu verwenden, zunehmend in der Diskussion, so etwa in Bericht der Federal Trade Commission, *Protecting Consumer Privacy In An Era Of Rapid Change: Recommendations For Businesses And Policymakers* von 2012.[210] Dieser Bericht zeigt aber auch die im Vergleich signifikanten Grenzen dieses Ansatzes, der ausschließlich eine freiwillige Selbstregulierung verlangt. Die Obama-Administration hatte zwar den Erlass verbindlicher Regeln vorgeschlagen,[211] doch ist es unwahrscheinlich ist, dass die jetzige Regierung, die bereits einige Datenschutzregulierungen der Obama-Zeit abgeschafft hat, diesem Vorschlag folgen wird. Der sektorale Ansatz zum Datenschutzrecht in den USA bedeutet, dass ein gänzlich neues *„Consumer Data Protection Law"* geschaffen werden müsste. Auch fehlt mit den ICOs nach europäischem Muster eine Behörde, die die Durchsetzung dieser Rechte überwachen würde. Der *Federal Trade Commission Act* (15 U.S.C. §§ 41-58) (FTC Act) kommt einem Verbraucherdatenschutzgesetz am nächsten und verbietet insbesondere ungerechtes oder täuschendes Verhalten.[212] Die FTC ist auch die primäre Verfolgungsbehörde für Verstöße gegen den *Children's Online Privacy Protection* Act (COPPA) (15 U.S.C. §§ 6501-6506) sowie die Selbstregulierungsprinzipien für *Behavioural Advertising*. Eine Generalkompetenz, algorithmisches Entscheiden zu regulieren, würde aber die Rolle der FTC fundamental ändern, insbesondere wenn diese Prinzipien nicht durch Selbstregierung (wie im Moment beim *Online Behavioural Advertising*) entstehen würden. So finden wir dann in der Diskussion vor allem Ideen, die hinter dem von der DSGVO

[207] Zur Verwendung von Datenschutz als Antidiskriminierungswerkzeug siehe Gellert et al. 2013, spezifisch Algorithmen betreffend siehe etwa Goodman/Flaxman 2016; Kuner et al. 2017; Borgesius/Poort 2017.
[208] Siehe z.B. Written Testimony of Frank Pasquale, Before the United States House of Representatives Committee on Energy and Commerce Subcommittee on Digital Commerce and Consumer Protection [https://bit.ly/2MiBgPI]; siehe auch Helveston 2015, S. 859; Jones 2015; Kearns 2018.
[209] UK Government: Data Science Ethical Framework, zuletzt besucht am 26.07.2018 [https://bit.ly/1sB2IEw].
[210] FTC Report (März 2012), zuletzt besucht am 26.07.2018 [https://bit.ly/1SHOpRB].
[211] So etwa im Brief von US-Präsident Barack Obama, der ein Appendix des White House, Consumer Data Privacy in a Networked World: A Framework for Protecting Privacy and Promoting Innovation in the Global Digital Economy (Feb. 23, 2012) verfasst hat.
[212] Siehe Hirsch 2014, S. 345.

Verlangten zurückbleiben und deshalb für unsere Diskussion irrelevant sind, oder die sich explizit an Europa orientieren.[213]

Trotzdem kann uns eine kurze vergleichende Analyse helfen, Antworten auf zwei relevante Fragen zu finden:

1) Zeigt die internationale Erfahrung, ob ein datenschutzrechtlicher Zugang vielleicht sogar ausreichend ist, um diskriminierende Algorithmen zu regulieren, so dass insoweit kein weiterer Regelungsbedarf besteht?
2) Gibt es mögliche Konflikte zwischen Datenschutz und der Kontrolle von ADM-Systemen?

Zu 1) finden sich in den USA Diskussionen, die dem Datenschutz nur eine minimale Schutzfunktion zutrauen.[214] Zum Teil sind diese durch ein verkürztes Verständnis des Datenschutzes erklärbar, der die Frage auf Zugangskontrolle und individuellen Konsens reduziert und nicht, wie dies gerade die DSGVO tut, die Legitimität der Datensammlung mit einer Prüfung der Fairness der Datenauswertung verbindet. Hier ist die Diskussion in Europa insofern sicher weiter.

Relevanter sind Sorgen, dass die Erfahrung mit dem bestehenden Datenschutzrecht zeigt, dass die Durchsetzung des Rechts den durchschnittlichen Benutzer überfordert, sowohl bezüglich des notwendigen Wissens als auch der finanziellen und zeitlichen Ressourcen. Gleichzeitig finden wir das gleiche Problem in der Durchsetzung von Diskriminierungsverboten.[215] Hier decken sich die Erfahrungen in den USA über das Durchsetzungsdefizit in beiden Bereichen durchaus mit denen in Europa. Andererseits wird aber die DSGVO den Datenschutzbehörden neue Durchsetzungsmittel geben, so dass auch hier das Bild gemischt ist. Wenn wir uns aber an den obigen Fall der diskriminierenden Versicherung in Großbritannien erinnern, so scheint hier zumindest die *Equality Commission* aktiver und effizienter als das ICO gewesen zu sein und die *Financial Conduct Authority* am stärksten involviert zu sein. Was man aus einem ersten Vergleich mit der Erfahrung sowohl im Datenschutz als auch im Diskriminierungsrecht in Europa und den USA sagen kann, ist, dass die Durchsetzung nicht ausschließlich den betroffenen Bürgern überlassen werden kann, sondern stattdessen Durchsetzungsorgane mit ausreichend Kompetenzen und Ressourcen benötigt werden. Dies bedeutet auch, dass Transparenz und ein „Recht auf Erklärung“ alleine wahrscheinlich nicht ausreichend sind.[216] Andererseits zeigt die Erfahrung in Großbritannien, dass Organisationen wie die *Financial Conduct Authority* zwar systematisch mit dem ICO zusammenarbeiten, aber mit einem Fairnessbegriff arbeiten, der zu unterschiedlich von dem des Datenschutzrechts ist, um eine völlige Übergabe der Verantwortung zu erlauben. Dies deckt sich auch mit der Position des britischen ICO zu einer möglichen Algorithmenaufsichtsbehörde. Einerseits weist das ICO in seinem Bericht zu *Big Data, Artificial Intelligence, Machine Learning And Data Protection* auf die positive Rolle des Datenschutzrechts und ihrer Behörde hin, und hier insbesondere auf Privacy Impact Assessment und eine zukünftige Zertifizierung.[217] Diese zentrale Rolle der Behörde wird auch in der Stellungnahme des ICO zur Konsultation des parlamentarischen Science and

[213] Siehe Webber/Car 2016, S. 76-88.
[214] So typisch etwa Dwork/Mulligan 2013.
[215] Siehe Milieu Ltd., „Comparative study on access to justice in gender equality and anti-discrimination law“ (Report for DG Justice, 2011), S. iii; Chopin/Germaine 2016, S. 81 ff.; Ellis/Watson 2012.
[216] So für datenschutzrechtliche Ansätze zur Algorithmenregulierung auch Edwards/Veale 2017.
[217] Information Commissioner’s Office: Big data, artificial intelligence, machine learning and data protection, zuletzt besucht am 26.07.2018 [https://bit.ly/2mF1kLj].

Technology Committees zu „Algorithms in decision-making“ betont.[218] Andererseits weist das ICO in derselben Stellungnahme korrekt darauf hin, dass nur ein kleiner Teil aller Algorithmen persönliche Daten verwendet und dass sozial schädliche, unter Umständen auch diskriminierende, Anwendungen denkbar sind, die keine persönlichen Daten verwenden und somit nicht unter den Anwendungsbereich der DSGVO fallen.[219]

Zumindest fraglich ist auch, ob das individuelle Recht auf Privatheit immer ausreichend ist, um negative Auswirkungen auf ganze Gruppen zu vermeiden.[220] Wir haben oben in der Diskussion zur Financial Conduct Authority bereits gesehen, wie im Versicherungswesen das Vorliegen von Behinderung ein rechtlich zulässiges Preisdiskriminierungskriterium sein kann und die negativen sozialen Folgen durch andere Maßnahmen wie etwa staatliche Zuschüsse abgemildert werden müssen. Die Labour-Opposition hat daher vorgeschlagen, bei der n Umsetzung der DSGVO den Data Protection Act 2018 hinsichtlich Art. 22 zu ergänzen. Der Vorschlag war, Sec 49 so zu ergänzen, dass einzelne Datensubjekte auch als Repräsentanten ihrer ganzen Gruppe handeln können.

Die neue Formulierung wäre gewesen:

49 A controller may not take a significant decision based solely on automated processing unless that decision is required or authorised by law.

(2) A decision is a “significant decision” for the purpose of this section if, in relation to a data subject, it —
(a) produces an adverse legal effect concerning the data subject, or

(b) significantly affects the data subject or *a group sharing a protected characteristic within the meaning of the Equality Act 2010 to which the data subject belongs*

Als Erklärung wurde als Beispiel die Wissenschaftlerin angegeben, die die „rassistischen“ Google-Werbealgorithmen untersuchte und feststellte, dass die Suche nach „schwarz klingenden“ Namen Werbung für Rechtshilfe in Strafrechtsfällen mit sich führte, nicht aber wenn sie nach „weiß klingenden“ Namen suchte, da sie, so Labour, als Forscherin in einer privilegierten Position ist und so durch diese Ergebnisse selber nicht „signifikant“ beeinträchtigt ist, aber das Recht haben sollte, im Namen ihrer Gruppe Korrekturen zu verlangen.

Der Vorschlag wurde zwar in der parlamentarischen Abstimmung abgelehnt, doch sagt auch das ICO, dass Datenschutz zwar ein wichtiger Teil der Algorithmenregulierung darstellen muss, aber gleichzeitig: „Wir sind uns bewusst, dass angesichts der Herausforderungen, die sich aus der Anwendung von Datenschutzgrundsätzen auf Big-Data-Analysen ergeben, ein anderer rechtlicher oder ordnungspolitischer Ansatz erforderlich ist. Wir akzeptieren jedoch nicht die Idee, dass der Datenschutz, wie er derzeit in der Gesetzgebung verankert ist, nicht in einem Big-Data-Kontext funktioniert.“[221] Im gleichen Dokument, und im gleichen Sinne,

[218] The Information Commissioner’s Office’s (ICO’s) response to the Science and Technology Committee’s call for evidence on algorithms in decision-making, zuletzt besucht am 26.07.2018 [https://bit.ly/2CEpKhV].
[219] Denkbar sind hier z.B. Algorithmen, die die Überflutungsgefahr für ganze Gegenden für Versicherungszwecke mit fehlerhaften Daten berechnen, oder ein Verkehrsalgorithmus, der durch die Ampelschaltung den Verkehr gezielt in ärmere Stadtgegenden umleitet.
[220] Siehe die Diskussion in Mantelero 2016.
[221] Information Commissioner’s Office: Big data, artificial intelligence, machine learning and data protection, zuletzt besucht am 26.07.2018 [https://bit.ly/2mF1kLj], eigene Übersetzung.

begrüßte das ICO die Einrichtung neuer „AI Ethics Boards“ als zusätzliches Regulierungsinstrumentarium, erwartet aber, dass diese eng mit dem ICO zusammenarbeiten werden und es nicht zu einer Duplikation von Funktionen führen soll.

Die internationale Diskussion legt damit nahe, dass die DSGVO einen erheblichen Beitrag zur Regulierung diskriminierender Algorithmen leisten kann, es aber noch weiteren Regelungsbedarf gibt.

Ein zweites Problem, ebenfalls insbesondere prominent in der US-Diskussion, sind mögliche Konflikte zwischen Datenschutz und Algorithmenregulierung und vor allem der Identifikation diskriminierender Algorithmen. So diskutieren Dwork und Mulligan stellvertretend für viele das Problem, dass z.B. Diskriminierung nach ethnischen Merkmalen zumindest schwerer aufzudecken ist, wenn kein expliziter Dateneintrag gemacht wurde, der die Ethnizität des Datensubjekts dokumentiert.[222] Dies ist in den USA auch im Rahmen der Diskussion zur positiven Diskriminierung und „Affirmative Action“ ein politisch sehr brisantes Thema: Ist es fairer, gar nicht erst nach Ethnizität oder Geschlecht zu fragen, oder müssen diese Kategorien, so umstritten sie auch sein können, weiter in der Datenerhebung verwendet werden, um im Audit die Neutralität der Ergebnisse beweisen zu können? Dwork und Mulligan argumentieren, dass Datenschutz hier als trojanisches Pferd verwendet wird, um eine politische Frage vorzuentscheiden und durch eine verpflichtende „farbenblinde“ Datenerhebung den Nachweis indirekter Diskriminierung zu erschweren. Dies ist kein allzu großes Problem, in Europa könnte etwa die Angabe der problematischen Merkmale dadurch gerechtfertigt werden, einer rechtlichen Verpflichtung nachzukommen.[223] Es ist aber durchaus richtig, dass Algorithmen-Audits und -Tests potenziell Datenschutzrisiken erzeugen können. So gibt die DSGVO dem Datensubjekt ein Informationsrecht, wie seine Daten durch einen Algorithmus bearbeitet wurden. Doch wie unsere Diskussion zur Enttarnung diskriminierender Algorithmen durch die Presse gezeigt hat, wird dies ihm normalerweise nicht helfen – was er braucht, sind die Entscheidungen, die für andere in ähnlichen Umständen wie seinen getroffen wurden, und diese sind ihm natürlich auch aus Datenschutzgründen nicht zugänglich.[224] So finden sich dann auch in der Tat Beispiele, in denen Firmen Datenschutz als Grund angaben, nicht mit der Ermittlung ihrer algorithmischen Methoden zu kooperieren. Dies wird auch eine lösbare Herausforderung für ein Algorithmen-Audit sein: Wie wir diskutiert haben, muss solch ein Audit regelmäßig auf die Trainingsdaten und/oder die Gesamtmenge der Entscheidungen Zugriff haben. Dies stellt natürlich auch ein potenzielles Datensicherheitsproblem dar, und die relevanten Methoden und Protokolle müssen sorgfältig entwickelt werden. Andererseits zeigt die amerikanische Erfahrung mit dem Fair Lending Audit, dass solche Analysen über Jahrzehnte vorfallfrei möglich sind.

Ein verwandtes Problem, das in der internationalen Diskussion, insbesondere in den USA, aufkam, sind andere rechtliche Gefahren des Testens durch Dritte. Hier bestehen Sorgen, dass vom Eigentümer nicht autorisierte Tests strafrechtliche Verstöße gegen den *Computer Fraud and Misuse Act* oder den *Digital Millennium Copyright Act* sein könnten. Dies kann dazu führen, dass *bona fide* Forscher, die etwa diskriminierendes Verhalten eines Algorithmus feststellen wollen, sich zumindest in einer rechtlichen Grauzone bewegen – ein

[222] Dwork/Mulligan 2013, S. 357 ff. für ähnliche Probleme in der DSGVO.
[223] Aber mit positivem Ergebnis bezüglich möglicher algorithmischer Audits siehe Goodman 2016, S. 493. Ein technischer Ansatz zur Lösung dieses Problems existiert auch, siehe Mancuhan/Clifton 2014, S. 211-238.
[224] Siehe dazu auch Henderson, Tristan, Does the GDPR Help or Hinder Fair Algorithmic Decision-Making? (Aug 21, 2017). Verfügbar unter SSRN: [https://ssrn.com/abstract=3140887].

Problem, das insbesondere auch aus der Cybersecurity-Forschergemeinde bekannt ist, wo es als Antwort anerkannte Regeln zur ethischen Offenlegung gefundener Schwachstellen gibt. Die American Civil Liberties Union (ACLU) hat nun Klage eingereicht, in der die Verfassungsmäßigkeit des Gesetzes gegen Computerbetrug und -missbrauch in Frage gestellt wird bzw. eine Klarstellung seiner Anwendbarkeit auf Forscher gegeben werden soll. Unter dem CFMA ist es illegal, auf einen Computer zuzugreifen, der „autorisierten Zugang" überschreitet. Diese Vorschrift verbietet potenziell Akademikern, Forschern und Journalisten das Testen auf Diskriminierung im Internet. Die Klage wurde im Juni 2016 beim District Court des District of Columbia als Sandvig v Sessions eingereicht.[225] Dies Problem ist nicht auf die USA beschränkt. In den oben diskutierten Enthüllungen möglicher Diskriminierung im Versicherungswesen hat eine der Parteien das zum Testen durchgeführte Antragstellen mit falschem Namen als einen möglichen Betrug bezeichnet und auch wegen Verleumdung mit Klage gedroht.

Auch das Urheberrecht wird manchmal als Barriere zur rechtskonformen Diskriminierungsanalyse genannt. In den USA wurde erst vor zwei Jahren eine entsprechende Ausnahme für Cybersecurity-Forscher geschaffen.[226] Eine ähnliche Regelung wird unter Umständen für die algorithmische Diskriminierungsanalyse notwendig sein. Für unsere Frage bedeutet diese Erfahrung zum einen, dass der rechtliche Rahmen des Algorithmen-Audits sowohl mit Datenschutz- als auch Urheberrecht koordiniert werden muss. Zum anderen zeigen sich potenzielle Gefahren, wenn Algorithmen auf Systemen analysiert werden, die sich in den USA befinden.

6.2.4 Kitemarks und Industrienormen

In der Diskussion der datenschutzrechtlichen Lösungen haben wir gesehen, wie unter anderem das ICO die Bedeutung der Zertifizierung unter Art. 43 DSGVO betont, und der Datenschutz hat sicher generell die Rolle von Zertifizierungen als Instrument der Algorithmenregulierung auch international in den Vordergrund gerückt.[227] Eine rechtsvergleichende Studie von Cavoukian und Chibba zeigt dabei eine verwirrende Vielzahl nationaler und internationaler Zertifizierungsprogramme mit oftmals variabler Qualität und Transparenz.[228] Die internationale Erfahrung mit Trustmarks ist hier eine zeitgerechte Warnung.[229] Die von der Europäischen Kommission finanzierte Studie zum Schutz der Privatsphäre der Bürger untersuchte bestehende Datenschutzsiegel und damit zusammenhängende Informationssiegel und siegelbasierte Zertifizierungssysteme in anderen Politikbereichen (einschließlich Telekommunikation, Bank- und Finanzwesen sowie Umweltvorschriften). Aus dieser vergleichenden Analyse konnten die Forscher eine Reihe von Kriterien für die Gestaltung und den Betrieb eines wirksamen Datenschutzsiegels ermitteln.[230] Eine detaillierte Diskussion dieser Studien geht über den Rahmen dieses Berichts hinaus; wir halten nur fest, dass es mittlerweile gesicherte Verfahren gibt, um die Transparenz, Effizienz und Akzeptanz solcher Siegel zu optimieren.

[225] Law Suit Sandvig v. Lynch – Complaint, zuletzt besucht am 26.07.2018 [https://bit.ly/2NtzDTW].
[226] US Federal Trade Commission, DMCA security research exemption for consumer devices, zuletzt besucht am 26.07.2018 [https://bit.ly/2Qj4twZ].
[227] Siehe Lachaud 2017.
[228] Cavoukian/Chibba 2018, S. 59-82.
[229] Balboni/Dragan 2018, S. 83-111.
[230] Siehe Rodrigues et al. 2013, S. 100-116.

Hier soll aber kurz auf einige der internationalen Initiativen verwiesen werden, die sich mit der Entwicklung zertifizierbarer Methodologien zum algorithmischen Audit und Best Practice Guidelines zur nichtdiskriminierenden KI beschäftigen.

Auf einem besonders hohen Abstraktionsniveau finden sich Dokumente wie die *Erklärung von Montreal für eine verantwortungsvolle Entwicklung der Künstlichen Intelligenz*, die am 3. November 2017 zum Abschluss des im Palais des congrès de Montréal stattfindenden Forums zur sozial verantwortlichen Entwicklung der KI angekündigt wurde.[231] Spezifisch das Problem des Maschinellen Lernens und der Algorithmen ist Gegenstand der am 16.5.2018 veröffentlichten und von Amnesty International und Access Now geschriebenen Toronto Declaration: *Protecting the rights to equality and non-discrimination in machine learning systems.*[232]

Bereits sehr viel spezifischer ist das EU-finanzierte Projekt „Werte und Ethik in der Innovation für verantwortungsvolle Technologie in Europa“ (VIRT-EU). Das Ziel des Projekts ist die Analyse und Abbildung der ethischen Praktiken von europäischen Hardware- und Softwareunternehmern, Maker- und Hacker-Spaces sowie Community-Innovatoren, um (1) zu verstehen, wie IoT-Innovatoren Ethik bei der Entwicklung zukünftiger Geräte umsetzen, zur (2) Schaffung eines neuen Rahmens für Datenschutz, ethische und soziale Folgenabschätzung (PESIA) und zur (3) Entwicklung von Instrumenten zur Unterstützung der ethischen Reflexion und Selbstbewertung als Teil des Designs. PESIA insbesondere enthält ein algorithmenspezifisches Äquivalent zur Privacy-Impact-Assessment-Methodologie des Datenschutzrechts.[233]

Ein US-Äquivalent findet sich im Algorithm Impact Assessment Tool, das vom AI Now Institute an der New Yorker Universität primär für den öffentlichen Sektor entwickelt wurde.[234] In Großbritannien wurde das oben zitierte Data Ethics Framework durch Nesta[235] zu einem Algorithmen-Risiko-Evaluierungsinstrument mit zehn ethischen Prinzipien und zugeordneten Risikokriterien zur Entwicklung von Algorithmen für die Verwaltung ausgebaut.[236]

Noch sehr viel spezifischer und detaillierter sind Standards, die von wissenschaftlichen Organisationen entwickelt werden. Das US-amerikanische Public Policy Council (USACM) der ACM (*Association for Computational Mechanics*) veröffentlichte eine Erklärung und eine Liste von sieben Prinzipien, die darauf abzielen, mögliche schädliche Verzerrungen bei

[231] The Montreal Declaration for a Responsible Development of Artificial Intelligence, zuletzt besucht am 26.07.2018: [https://bit.ly/2Nu3ag9].
[232] The Toronto Declaration: Protecting the rights to equality and non-discrimination in machine learning systems, zuletzt besucht am 26.07.2018: [https://bit.ly/2McCD2o].
[233] Vgl. Webseite des Ethos Labs der IT University of Copenhagen: The VIRT-EU Project, zuletzt besucht am 26.07.2018: [https://bit.ly/2Qilkjy].
[234] Vgl. Webseite des AI Now Institutes: Algorithmic Impact Assessment (April 2018), zuletzt besucht am 26.07.2018: [https://bit.ly/2p1t1hX].
[235] Die *National Endowment für Wissenschaft, Technologie und Kunst* (Nesta) ist heute eine unabhängige Stiftung, ursprünglich aber ein durch Parlamentsbeschluss eingerichteter „non-departmental public body“, unabhängig von der Regierung, aber mit einem Minister als letztendlich Verantwortlichem, mit besonders engen Verbindungen zu Regierung und Verwaltung.
[236] Vgl. Webseite von Nesta: 10 principles for public sector use of algorithmic decision making, zuletzt besucht am 26.07.2018: [https://bit.ly/2QoaOrg].

algorithmischen Lösungen zu beseitigen. Diese Bemühungen wurden von der *Algorithmic Accountability Working Group* des USACM initiiert.[237]

Prinzipien für algorithmische Transparenz und Verantwortlichkeit sind folgende:

1. **Awareness:** Eigentümer, Designer, Benutzer und andere Beteiligte von Analysesystemen sollten sich der möglichen Vorurteile bewusst sein, die mit ihrer Gestaltung, Implementierung und Verwendung sowie dem möglichen Schaden, den Voreingenommenheit Einzelpersonen und der Gesellschaft verursachen kann, verbunden sind.

2. **Zugang und Rechtsbehelfe:** Die Regulierungsbehörden sollten die Einführung von Mechanismen fördern, die Befragungen und Rechtsbehelfe für Einzelpersonen und Gruppen ermöglichen, die von algorithmisch fundierten Entscheidungen betroffen sind.

3. **Rechenschaftspflicht:** Institutionen sollten für Entscheidungen verantwortlich gemacht werden, die durch die von ihnen verwendeten Algorithmen getroffen werden, auch wenn es nicht möglich ist, im Detail zu erklären, wie die Algorithmen ihre Ergebnisse produzieren.

4. **Erläuterung:** Systeme und Institutionen, die algorithmische Entscheidungen treffen, werden ermutigt, Erläuterungen sowohl zu den vom Algorithmus verfolgten Verfahren als auch zu den spezifischen getroffenen Entscheidungen zu geben. Dies ist besonders wichtig in öffentlichen politischen Kontexten.

5. **Datenprovenienz:** Eine Beschreibung der Art und Weise, in der die Trainingsdaten gesammelt wurden, sollte von den Erstellern der Algorithmen beibehalten werden, begleitet von einer Untersuchung der möglichen Verzerrungen, die durch den menschlichen oder algorithmischen Datenerfassungsprozess induziert werden. Die öffentliche Überprüfung der Daten bietet maximale Möglichkeiten für Korrekturen. Bedenken hinsichtlich des Datenschutzes, des Schutzes von Geschäftsgeheimnissen oder der Offenlegung von Analysen, die böswilligen Akteuren erlauben könnten, das System zu manipulieren, können jedoch den Zugang auf qualifizierte und autorisierte Personen einschränken.

6. **Überprüfbarkeit:** Modelle, Algorithmen, Daten und Entscheidungen sollten aufgezeichnet werden, damit sie in Fällen, in denen ein Schaden vermutet wird, überprüft werden können.

7. **Validierung und Testen:** Institutionen sollten strenge Methoden anwenden, um ihre Modelle zu validieren und diese Methoden und Ergebnisse zu dokumentieren. Insbesondere sollten sie routinemäßig Tests durchführen, um zu beurteilen und festzustellen, ob das Modell diskriminierende Schäden verursacht. Die Institutionen werden ermutigt, die Ergebnisse solcher Tests öffentlich zu machen.

Ein etwas weiter im Detail ausgearbeiteter Vorschlag ist das Weißbuch des Global Future Council on Human Rights des World Economic Forum *How to Prevent Discriminatory Outcomes in Machine Learning*. Es bietet Entwicklern einen Rahmen, um Diskriminierung bei der Entwicklung des Maschinellen Lernens zu verhindern. Das Dokument wurde nach einer langen Konsultationsphase erstellt und basiert auf Recherchen und Interviews mit Branchenexperten, Wissenschaftlern und Menschenrechtsexperten – in dieser Hinsicht bestätigt es den oben angeführten Punkt, dass auch die Industrie die Notwendigkeit von rechtlichen und ethischen Normen akzeptiert. Die Methodologie lehnt sich auf juristischer Seite stark an das internationale Menschenrecht an. Die Empfehlung für Entwickler und alle

[237] ACM US Technology Policy Committee: Statement on Algorithmic Transparency and Accountability, 12.01.2017, zuletzt besucht am 16.07.2018: [https://bit.ly/2iopli5].

Unternehmen, die Maschinelles Lernen nutzen möchten, lautet, Nichtdiskriminierung zu priorisieren, indem ein auf vier Leitprinzipien basierender Rahmen festgelegt wird: aktive Einbeziehung, Gerechtigkeit, Recht auf Erklärung und Zugang zu Wiedergutmachung.[238]

Die IEEE Standards Association, ein weltweit anerkanntes Gremium für Standards innerhalb der IEEE, entwickelt Konsensstandards durch einen offenen Prozess, der die Industrie einbezieht und eine breite Interessengemeinschaft zusammenbringt. Die IEEE-Standards legen Spezifikationen und Best Practices auf der Grundlage des aktuellen wissenschaftlichen und technologischen Wissens fest. Die IEEE-SA hat ein Portfolio von über 1.250 aktiven Standards und mehr als 650 weitere Standards in der Entwicklung. Die IEEE *Global Initiative on Ethics of Autonomous and Intelligent Systems* beschäftigt sich mit der kollaborativen Entwicklung von offenen Standards zu Fragen des ethischen und rechtskonformen Designs sowie einem Modellprozess zur Behandlung ethischer Bedenken während des Systemdesigns (IEEE P7000)[239] und arbeitet aktiv an einer Reihe von verwandten Standards in diesem Bereich, einschließlich Transparenz[240], Datenschutz, Diskriminierung und Governance.[241]

Während die IEEE-Projekte noch in der Entwicklung sind, hat die Statistikerin und Autorin des einflussreichen Buchs *Weapons of Math Destruction* Cathy O'Neill ein Gütesiegel für Algorithmen entwickelt, das ihre Genauigkeit, Unvoreingenommenheit und Fairness evaluiert.[242] Die Zertifizierung findet durch ihre eigene Firma O'Neil Risk Consulting und Algorithmic Auditing (ORCAA) statt; das erste so zertifizierte Unternehmen ist das Vermietungsunternehmen Rentlogic.

Diese erste Übersicht stellt keinen Anspruch auf Vollständigkeit. Sie zeigt aber, dass in verschiedenen Graden der Granularität bereits Best Practice Guides, Standards und Gütesiegel existieren, so dass ein etwaiges neues Prüfinstitut auf ein getestetes, international zunehmend einflussreiches und wissenschaftlich robustes Rahmenwerk zurückgreifen könnte und dann entweder im Licht der abstrakteren Vorschläge ein eigenes, konkretes Prüfungs- und Zertifizierungsverfahren entwickeln kann, oder – und dies ist der Vorschlag der UK-Regierung für eine solche Institution – als „Prüfer der Prüfer" den besten existierenden Verfahren öffentliche Anerkennung gibt.

6.2.5 Wettbewerbsrecht

Algorithmische Preisdiskriminierung ist auch ein Thema für das Wettbewerbsrecht.[243] Insbesondere in den USA finden wir den Gedanken, dass die Regulierung algorithmischer Entscheidungsfindung besser den Märkten und der Selbstregulierung überlassen bleiben sollte[244]. Dies setzt aber effiziente Märkte mit geringen Transaktionskosten voraus, was

[238] White Paper des World Economic Forums: How to Prevent Discriminatory Outcomes in Machine Learning, März 2018, Downlaod am 26. Juli 2018 [https://bit.ly/2p9WjdK].
[239] Webseite von IEEE Project: 7000 - Model Process for Addressing Ethical Concerns During System Design, zuletzt besucht am 26.07.2018 [https://bit.ly/2oSnLgi].
[240] Webseite von IEEE Project: 7001 - Transparency of Autonomous Systems, zuletzt besucht am 26.07.2018 [https://bit.ly/2x7WZo7].
[241] Webseite von The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, zuletzt besucht am 26.07.2018 [https://bit.ly/1USPRDW].
[242] Webseite von O'Neil Risk Consulting & Algorithmic Auditing, zuletzt besucht am 26.07.2018 [www.oneilrisk.com].
[243] So etwa Vestager 2017; Ohlhausen 2017; siehe aus vergleichender Perspektive Gosselin 2017; siehe auch Ezrachi/Stucke 2017; Harrington 2017; Mehra 2016, S. 1323-1375.
[244] Okuliar/Kamenir 2017.

durch intransparente Algorithmen in Frage gestellt wird.[245] Diese Diskussion ist international eine der am weitesten entwickelten. Neben Großbritannien und den USA finden wir sie auch in Ländern wie Russland,[246] Frankreich und Australien. In diesem Bereich bestehen auch erste Ansätze zu einer internationalen Regulierung – auf EU-Ebene gibt es seit 2017 eine Analyse der Preissetzung in Onlinemärkten, die insbesondere auch Pricing-Software behandelt. Die OECD nahm sich des Themas 2016 an[247] und wenngleich der OECD-Konsensus bislang lautet, dass Preisdiskriminierung nur in Ausnahmefällen ein wettbewerbsrechtliches Problem ist, werden gerade algorithmische Kollisionen als eine mögliche Ausnahme genannt. Eine detaillierte Analyse der wettbewerbsrechtlichen Probleme der Algorithmen geht über den Rahmen dieser Studie hinaus. Bemerkt werden soll nur, dass die internationale Debatte hier ein Problem aufgegriffen hat, das als Nebenwirkung bedeuten kann, dass transparenz- und marktbasierte Lösungen zum Problem diskriminierender Algorithmen zu kurz greifen werden.

6.2.6 Verbraucherschutzrecht und Verbraucherpanel

In diesem Zusammenhang ist in den USA z.B. der Vorschlag von *Consumer Subject Review Boards* zu nennen, nach dem Vorbild des Belmont Reports von 1987 zur Ethik der Studien an menschlichen Versuchsobjekten.[248] Diese würden ein System von betriebsinternen Ethikgruppen mit Verbraucherbeteiligung mit sich führen, in den USA mit der FTC als Koordinator. Diese Form der Selbst- oder Koregulierung wird auch von der britischen Regierung befürwortet und ist wahrscheinlich eine der Rollen, die das neue nationale Datenethikinstitut wahrnehmen soll. Hintergrund sind hier insbesondere Vorfälle wie Facebooks Versuch, mit Hilfe von Wissenschaftlern der Harvard University Einsichten in die emotionale Auswirkung von News zu erlangen, um Facebooks Newsfeed-Algorithmen entsprechend zu optimieren, und dazu systematisch seine Benutzer ohne deren Wissen manipulierte.[249] Der Vorteil dieses Ansatzes ist die Flexibilität und mögliche Reichweite – die Facebook-Forschung war nach US-Recht legal, führte aber zu einem öffentlichen Aufschrei, einem Reputationsverlust und einer Entschuldigung. Ein Ethics Advisory Panel hätte dies verhindern können. Es bleibt aber in beiden Ländern unklar, welche Art (und Größe) von Unternehmen solch ein Panel einsetzen soll, wie hoch die Kosten sein werden und welche Konsequenzen daraus folgen werden, wenn ein Unternehmen die Ratschläge des Panels ignoriert.

6.2.7 Sui-generis-Ansatz mit neuer Aufsichtsbehörde

Die letzte Gruppe von Lösungsansätzen ist vor allem in Großbritannien die von den aktuellen Rechtsreformen am weitesten vorangeschrittene. Eine neue Aufsichtsbehörde mit Verantwortung für das gesamte Feld der *„Algorithmic Governance“* soll sich dort mit allen Fragen der AI-Regulierung beschäftigen.

Ein dementsprechender Vorschlag war zum ersten Mal 2016 in einem Bericht des House of Commons Science and Technology Committees gemacht worden.[250] Die britische Regierung sollte die neue Kommission am Alan Turing Institute gründen, einem Verband

[245] U.S. Department of Justice 2017.
[246] Algorithms and Collusion - Note by the Russian Federation [https://bit.ly/2qibl2g].
[247] OECD Background Note by the Secretariat, Algorithms and Collusion, Roundtable on Algorithms and Collusion, DAF/COMP (2017) 4, June 2017.
[248] Calo 2013, S. 97.
[249] Kramer et al. 2014, S. 8788-8790.
[250] Webseite des UK-Parlaments, zuletzt besucht am 26. Juli 2018: [https://bit.ly/2QjliqL].

aus den (damals fünf, jetzt neun) forschungsstärksten Universitäten im Bereich Datenwissenschaft, mit einem physischen Zentrum in London.

Die Kommission sollte Mitglieder aus den verschiedensten Bereichen umfassen, einschließlich Experten der Rechtswissenschaften, der Sozialwissenschaften und der Philosophie, Informatiker, Naturwissenschaftler, Mathematiker und Ingenieure sowie Vertreter von Industrie, Nichtregierungsorganisationen und der Öffentlichkeit.

„[Die Kommission] sollte sich auf die Festlegung von Grundsätzen für die Entwicklung und *Anwendung von KI-Techniken konzentrieren und die Regierung bei der Festlegung von Beschränkungen für deren Fortschreiten beraten. Sie muss eng mit der Arbeit des Rates für Datenethik koordiniert werden, den die Regierung derzeit aufstellt.“*

Das Brexit-Referendum führte dazu, dass dieser Vorschlag erst einmal nicht umgesetzt wurde und nur eine Data Ethics Group am Turing Institute mit Repräsentanten der beteiligten Universitäten eingesetzt wurde, die seitdem insbesondere auch Machine-Learning-Projekte im öffentlichen Sektor, einschließlich der Polizei, aus ethischer Perspektive unterstützt und berät.

Eine parteiübergreifende parlamentarische Expertengruppe (APPG) zur KI mit Mitgliedern des House of Lords und des House of Commons wurde im Januar 2017 mit dem Ziel eingesetzt, den Einfluss und die Auswirkungen Künstlicher Intelligenz, einschließlich des Maschinellen Lernens, zu untersuchen. Sie legte im Januar 2017 einen Bericht mit sieben Kernempfehlungen und dazugehörigen Berichten vor.[251] Angehört wurden 1309 Experten und Stakeholder aus der Industrie. Drei der Berichte befassen sich mit Fragen der Regulierung: Theme Report 2: Ethics and Legal in AI: Decision Making and Moral Issues; Theme Report 3 mit dem Titel „Ethics and Legal in AI: Data Capitalism“ und Theme Report 5: Governance, Social and Organisational Perspective for AI. Vorgeschlagen als institutionelle Antworten auf KI wurden eine neue Regierungsbehörde für KI, ein neuer von der Industrie geleiteter KI-Rat, ein neuer GovTech-Katalysator, ein neues Future-Sectors-Team und, für unsere Frage am wichtigsten, ein neues Zentrum für Datenethik und -innovation, das in enger Koordination mit dem Information Commissioner’s Office alle anderen Gruppen zu Fragen von Recht und Ethik beraten soll. Wie bereits angedeutet, zeigte eine Analyse der Beiträge der Industrievertreter weitgehend Zustimmung, dass Regulierung zum Erhalt des öffentlichen Vertrauens notwendig und wünschenswert sei – und nahm sich selber in die Pflicht, mehr zur Gestaltung dieses Rechtsrahmens beizutragen.[252]

Parallel zu diesem Projekt beauftragte die Regierung die Royal Society, ein Politikprojekt zum Maschinellen Lernen durchzuführen. Damit sollten das Potenzial des Maschinellen Lernens in den nächsten fünf bis zehn Jahren und die Hindernisse für die Realisierung dieses Potenzials untersucht werden. Bei der Durchführung der Untersuchung beschäftigte sich das Projekt mit Zielgruppen in Politikkreisen, der Industrie, der Wissenschaft und der Öffentlichkeit, um das Bewusstsein für Maschinelles Lernen zu wecken, Ansichten der Öffentlichkeit zu verstehen und zur öffentlichen Debatte über diese Technologie beizutragen. Fragen der rechtlichen Ausgestaltung gehörten mit zur Aufgabe.[253] Flankierend organisierte auch die Law Society for England and Wales Roundtables zu KI und zum

[251] All-Party Parliamentary Group on Artificial Intelligence, Big Innovation Centre 2017: Findings 2017, Download am 26. Juli 2018: [https://bit.ly/2QgplFj].
[252] Siehe z.B. die Intervention von Ben Taylor (CEO at RainBird Technologies), Download am 26. Juli 2018 [https://bit.ly/2CMmmBY].
[253] Webseite der Royal Society, zuletzt besucht am 26. Juli 2018 [https://bit.ly/2Mhwwtt].

Rechtsstaatsprinzip.[254] 2017 begann der Ausschuss für Künstliche Intelligenz des House of Lords seine eigene Untersuchung, zu der mündliche Aussagen von über 200 Experten gesammelt wurden. Er legte seinen Bericht *AI in the UK: ready, willing and able?* am 16.4.2018 vor.

In diesem lehnt es zwar eine umfassende KI-spezifische Regelung ab und argumentiert für eine Stärkung der bestehenden sektorspezifischen Aufsichtsbehörden. Doch soll das neue Zentrum für Datenethik und -innovation, für das die Regierung im Haushalt für 2017 bereits Mittel bereitgestellt hat, damit beauftragt werden die Lücken aufzeigen, die dort bestehen, wo die bisherigen Vorschriften möglicherweise nicht ausreichen. Ein Kernbestand weithin anerkannter ethischer Prinzipien, auf die sich Unternehmen und Organisationen berufen, die KI einsetzen, soll mit substanziellen Beiträgen des Zentrums für Datenethik und -innovation, des AI Councils und des Alan Turing Institutes entwickelt werden.

So argumentiert das House of Lords in Sec 419 und 420:

„Viele Organisationen bereiten ihre eigenen ethischen Verhaltenskodizes für den Einsatz von AI vor. Diese Arbeit ist zu würdigen, aber es ist klar, dass es an einer breiteren Sensibilisierung und Koordination fehlt, wo die Regierung helfen könnte. Konsequente und allgemein anerkannte ethische Leitlinien, denen sich Unternehmen und Organisationen, die AI einsetzen, anschließen, wären eine willkommene Entwicklung."

Wir empfehlen, dass ein sektorübergreifender ethischer Verhaltenskodex („AI-Code"), der für die Umsetzung in öffentlichen und privaten Organisationen, die AI entwickeln oder einführen, geeignet ist, vom *Centre for Data Ethics and Innovation Consultation* erstellt und gefördert wird. Dieser soll mit Beiträgen des AI Councils und des Alan Turing Institutes und mit einer gewissen Dringlichkeit behandelt werden. In einigen Fällen müssen branchenspezifische Variationen erstellt werden, die eine ähnliche Sprache verwenden. Ein solcher Kodex sollte die Notwendigkeit beinhalten, die Einrichtung von Ethikbeiräten in Unternehmen oder Organisationen in Betracht zu ziehen, die AI bei ihrer Arbeit entwickeln oder einsetzen. Mit der Zeit soll der KI-Code die Grundlage für gesetzliche Regelungen bilden, wenn dies für notwendig erachtet wird.

Dabei stellt das *House of Lords* insbesondere heraus, dass sich Großbritannien auch weltweit als Vorreiter in der ethischen und rechtskonformen KI etablieren und dies zu einem „Unique Selling Point" machen will, so dass die ethische Algorithmenentwicklung auch zu einem Wettbewerbsvorteil für die Industrie führt.

Mögliche Aufgaben für das Datenethik-Institut, das durch Gesetz als regierungsunabhängige Behörde eingerichtet werden wird, beinhalten dabei:

- Best-Practice-Richtlinien mit der Wirtschaft zu entwickeln
- in Zusammenarbeit mit dem ICO ein Prüfungszeichen und relevante Tests zu entwickeln
- als „trusted third party" eine zentrale Sammelstelle für Algorithmen zu sein, die dann im Streitfall selektiv analysiert werden können
- ein System der „postmarketing surveillance" für Algorithmen zu etablieren

[254] Bingham Center for the Rule of Law, zuletzt besucht am 26. Juli 2018 [https://bit.ly/2oUgkVT].

Der Bericht fordert die Regierung auch dazu auf, dringende Schritte zu unternehmen, um die Schaffung von maßgeblichen Instrumenten und Systemen für die Prüfung und den Test von Trainingsdatensätzen zu unterstützen, um sicherzustellen, dass sie repräsentativ für verschiedene Bevölkerungsgruppen sind und dass sie beim Training von KI-Systemen das Risiko diskriminierender Entscheidungen minimalisieren. Er empfiehlt eine öffentlich finanzierte „Herausforderung", um Anreize für die Entwicklung von Technologien zu schaffen, die KIs prüfen und befragen können.

„Das Zentrum für Datenethik und -innovation sollte in Absprache mit dem Alan-Turing-Institut, dem Institut für Elektro- und Elektronikingenieure, dem British Standards Institute und anderen Fachgremien Leitlinien zu der Anforderung an die Verständlichkeit von KI-Systemen erarbeiten", fügt der Ausschuss hinzu. „Der KI-Entwicklungssektor sollte sich bemühen, solche Leitlinien zu übernehmen und sich unter der Schirmherrschaft des AI-Rates auf die für die Sektoren, in denen er tätig ist, relevanten Normen zu einigen."[255]

Für unseren Bericht ist an diesem Vorschlag vielleicht am wichtigsten, dass er als Ergebnis von drei Jahren intensiver Zusammenarbeit mit Industrie und Universitäten entstand und damit einen guten Einblick eröffnet, was von der Wirtschaft als tragbar empfunden wird. Hier ist nicht nur das generelle Interesse an Regulierung zu nennen, sondern gerade auch der Gedanke, dass ethisch und rechtlich verantwortliche KI ein massiver Wettbewerbsvorteil sein kann.

6.2.8 Europäische Initiativen

Seit dem Beginn der Arbeit an dieser Studie haben auch die Europäische Union und der Europarat in einer Reihe von Initiativen begonnen, eine Antwort auf die Frage der Regulierung der KI und Robotik zu finden. Bislang hat indes noch keines dieser Projekte zu neuen Gesetzen geführt (wenn man die Datenschutz-Grundverordnung und ihre Vorschriften zum automatischen Entscheiden einmal außen vor lässt) oder auch nur zu einem hinreichend expliziten Regulierungsvorschlag. Trotzdem seien die wichtigsten Projekte hier kurz angerissen, um sicherzustellen, dass unsere Vorschläge zumindest nicht mit absehbaren Initiativen in Konflikt stehen werden.

Ein erster Meilenstein war die *Entschließung* des Europäischen Parlaments vom 16. Februar 2017 mit Empfehlungen an die Kommission zu zivilrechtlichen Regelungen im Bereich Robotik (2015/2103(INL)).[256] Diese Entschließung war durch eine Studie des *European Parliament's Legal Affairs Committee* zu *European Civil Law Rules on Robotics* vorbereitet worden.[257]

[255] Open consultation: Centre for Data Ethics and Innovation Consultation, published 13.06.2018, zuletzt besucht am 8. August 2018 [https://bit.ly/2Otzfm6].
[256] Entschließung des Europäischen Parlaments vom 16. Februar 2017 mit Empfehlungen an die Kommission zu zivilrechtlichen Regelungen im Bereich Robotik (2015/2103(INL)), zuletzt besucht am 26. Juli 2018 [https://bit.ly/2MQ6YIs].
[257] European Parliament: DC for Internal Policies, 2016: European Civil Law Rules in Robotics: Download am 26. Juli 2018 [https://bit.ly/2iIgePS].

Die *Entschließung* bezog sich zwar explizit ausschließlich auf „cyber-physische Systeme", das heißt „verkörperte KI",[258] doch verweisen mehrere Vorschläge auf Maschinelles Lernen und generelle Probleme der KI. So besagt etwa Absatz H der Einleitung:

> *„dass Maschinelles Lernen der Gesellschaft durch eine deutliche Verbesserung der Datenanalysefähigkeit enorme wirtschaftliche und innovationsbezogene Vorteile bietet, aber auch Herausforderungen im Zusammenhang mit der Durchsetzung von Nichtdiskriminierung, ordnungsgemäßen Verfahren, Transparenz und der Verständlichkeit der Entscheidungsfindung mit sich bringt."*

Das Thema der Nichtdiskriminierung und des Maschinellen Lernens wird dann in den Empfehlungen aufgegriffen. Dabei vertritt die *Entschließung* die Ansicht, dass der bestehende EU-Rechtsrahmen modernisiert und ergänzt werden muss. Insbesondere relevant für unsere Studie sind die Abschnitte 12 und 13:

> *„12. betont den Grundsatz der Transparenz, wonach es jederzeit möglich sein muss, die Gründe für jede mithilfe der KI getroffene Entscheidung anzugeben, die sich wesentlich auf das Leben einer oder mehrerer Personen auswirken kann; ist der Auffassung, dass es jederzeit möglich sein muss, die Berechnungen von KI-Systemen zurück in eine für den Menschen verständliche Form zu überführen; […]"*

> *„13. weist darauf hin, dass der ethische Leitrahmen auf den Grundsätzen der Benefizienz, der Schadensverhütung, der Autonomie und der Gerechtigkeit sowie auf den in Artikel 2 des Vertrags über die Europäische Union und in der Charta der Grundrechte verankerten Grundsätzen und Werten beruhen sollte, wie z.B. Menschenwürde, Gleichheit, Gerechtigkeit und Fairness, Nichtdiskriminierung, Einwilligung nach Aufklärung, Privat- und Familienleben und Datenschutz sowie auf anderen dem Unionsrecht zugrundeliegenden Grundsätzen und Werten, wie Nichtstigmatisierung, Transparenz, Autonomie und individuelle Verantwortung und soziale Verantwortung, und auf bestehenden ethischen Praktiken und Regelwerken."*

Insbesondere Abschnitt 13 ist wahrscheinlich für „reine" KI, wie sie in den Szenarien dieser Studie verwendet wird, sogar noch relevanter als für cyber-physische Systeme. Den Gedanken des „interpretierbaren Algorithmus" und der Erklärbarkeit maschinell getroffener Entscheidungen spiegeln Art. 18 und 22 der Datenschutz-Grundverordnung wider, gehen aber über diese hinaus und sind eindeutiger in ihrer Anforderung. Während „Erklärbarkeit" im Datenschutzrecht in erster Linie dem Datensubjekt helfen soll, informierte Entscheidungen zu treffen, ist hier Erklärbarkeit eng an das Haftungsrecht angebunden und soll a posteriori eine Zurechenbarkeit von Verantwortung für falsche Entscheidungen ermöglichen. Dies bringt diese Ansätze sehr viel näher an das Ex-Post-Testen, das in unserer Studie im Vordergrund steht. Interpretierbarkeit der algorithmischen Modelle und statistisches Testen der Ergebnisse des Algorithmus sind damit zwei komplementäre Ansätze. Interpretierbare Algorithmen haben insbesondere im Rahmen der Diskussion um algorithmische Transparenz in den letzten Jahren großes Interesse geweckt;[259] sollten die in

[258] Vgl. Allgemeine Grundsätze 1. Zum Problem dieser Einschränkung auf physische Systeme siehe Schafer: „Closing Pandora's box?: The EU proposal on the regulation of robots." *Pandora's Box* 2016, S. 55.

[259] Siehe etwa Bach, Sebastian, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller und Wojciech Samek. „On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation" *PloS one* 10, no. 7 (2015): e0130140. Fong, R.C. und Vedaldi, A., 2017. Interpretable explanations of black boxes by meaningful perturbation. *arXiv preprint*

der Entschließung angedachten Erklärungspflichten angenommen werden, ist zu erwarten, dass dieser Forschungsansatz noch intensiver verfolgt werden wird – mit Lösungen, die mit den von uns vorgeschlagenen kompatibel sind.

Letztlich regt die *Entschließung* auch die Einrichtung einer europäischen Agentur für Robotik und Künstliche Intelligenz an, die sich auch gerade mit rechtlichen und ethischen Fragen der KI auseinandersetzen soll. Dabei soll der Schwerpunkt aber auf grenzüberschreitenden physischen Systemen liegen, womit die Bedrohungsszenarien von den in dieser Studie behandelten sehr verschieden sind. Zu denken ist insbesondere an „Entscheidungen", die ein Algorithmus in einem intelligenten Fahrzeug trifft. Trotz dieser Unterschiede sind die notwendigen technischen und mathematischen Lösungen für etwa das Testen der Korrektheit von Entscheidungen zum Teil zumindest gleich, so dass von einer solchen Agentur, die in der Fassung der *Entschließung* ausschließlich beratende und forschungskoordinierende Aufgaben wahrnimmt, wichtige Anregungen auch für nationale und sektorspezifische Ansätze wie den in dieser Studie vorgeschlagenen kommen werden. Konflikte oder die Duplizierung von Bemühungen würde es aber nicht geben, dazu sind sowohl der Aufgabenbereich als auch die Befugnisse dieser Agentur von dem hier Vorgeschlagenen zu unterschiedlich und letztlich komplementär.

Die Union hat einige der Anregungen bereits umgesetzt, insbesondere durch die Berufung einer *Expert Group on liability and new technologies* und einer *High-Level Expert Group on Artificial Intelligence*, die sich auch mit Fragen der ethischen und rechtlichen Ausgestaltung der KI befassen wird. Insbesondere die High Level Expert Group wird sich nicht nur, oder auch nur schwerpunktmäßig, mit cyber-physischer KI befassen. Ihr Aufgabenbereich ist sehr weit und soll insbesondere „KI-Ethikleitlinien zu Themen wie Fairness, Sicherheit, Transparenz, Zukunft der Arbeit, Demokratie und allgemeiner die Auswirkungen auf die Anwendung der Charta der Grundrechte, einschließlich Schutz der Privatsphäre und der personenbezogenen Daten, Würde, Verbraucherschutz und Nichtdiskriminierung" behandeln. Außer Wissenschaftlern und Industrie sind auch Initiativen von NGOs wie AlgorithmWatch in dieser Expertengruppe vertreten. Der besonders weite Aufgabenbereich, zusammen mit der Größe (60 Mitglieder) und Heterogenität dieser Gruppe, macht es wahrscheinlich, dass sämtliche Vorschläge detailarm und wohl auch nicht in absehbarer Zeit vorgelegt werden. Zudem sollen die Leitlinien auf der Arbeit der *Europäischen Gruppe für Ethik in den Naturwissenschaften und neuen Technologien* (die Europäische Gruppe für Ethik in den Wissenschaften und neuen Technologien (EGE) ist ein unabhängiges Beratungsgremium des Präsidenten der Europäischen Kommission) und *der EU Agentur für Grundrechte* aufbauen, die zurzeit eine Bewertung der aktuellen Herausforderungen an Hersteller und Nutzer neuer Technologien in Bezug auf die Einhaltung der Grundrechte zu bewältigen haben (Projekt „Big Data und Grundrechte").

Parallel zu diesen Bemühungen hat *Atomium – European Institute for Science, Media and Democracy* eine Expertengruppe, AI4People, eingesetzt, die im November 2018 der Kommission eine „Ethical Roadmap for AI" vorlegen wird.[260] Es ist wahrscheinlich, dass auch diese Roadmap nach einem Zertifizierungssystem für Algorithmen fragen wird, das insbesondere auch Fragen der Diskriminierung beinhalten soll.

Schließlich ist noch die *Study On The Human Rights Dimensions Of Automated Data Processing Techniques (In Particular Algorithms) And Possible Regulatory Implications* des

---

arXiv:1704.03296; Samek, W., Wiegand, T. und Müller, K.R., 2017. Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *arXiv preprint arXiv:1708.08296*.

[260] http://www.eismd.eu/ai4people/.

Europarats zu nennen.[261] Im Fokus der Studie stehen algorithmische Entscheidungen durch „rein digitale“ Programme, und sie ist daher dem Gegenstand unserer Studie einerseits besonders nahe, andererseits aber auch sehr viel weitreichender, mit einem besonderen Schwerpunkt auf Anwendungen durch die öffentliche Hand und in Gerichtsverfahren. Folgende Abschnitte sind von direkter Relevanz für unsere Studie: Sektion 2 behandelt den Datenschutz, Sektion 5 die Auswirkungen auf Artikel 13 der Europäischen Menschenrechtskonvention – Recht auf wirksame Beschwerde – und Sektion 6 Fragen der Diskriminierung.

Die Studie analysiert drei übergreifende Problemstellungen: Transparenz, Zurechenbarkeit/Verantwortlichkeit und ethische Rahmenbedingungen für eine bessere Risikobewertung. Für Letztere wird insbesondere auf die auch in diesem Bericht diskutierten IEEE-Standards verwiesen (IEEE P7000: Model Process for Addressing Ethical Concerns During System Design; IEEE P7001: Transparency of Autonomous Systems; IEEE P7002: Data Privacy Process; IEEE P7003: Algorithmic Bias Considerations).

Die Studie kommt zu Empfehlungen, von denen drei direkt für uns relevant sind:

Empfehlung 5 verlangt Folgendes:

> *„Zertifizierungs- und Prüfmechanismen für automatisierte Datenverarbeitungstechniken wie Algorithmen sollen entwickelt werden, um die Einhaltung der Menschenrechte sicherzustellen. Öffentliche Einrichtungen und nichtstaatliche Akteure sollten die Weiterentwicklung der Menschenrechte durch Design und ethische Ansätze fördern und die Annahme stärkerer Risikobewertungsansätze bei der Entwicklung von Software fördern.“*

Empfehlung 6 fordert:

> *„Öffentliche Stellen sollten mit ihren eigenen Regulierungsbehörden (Versicherungen, Kreditauskunfteien, Banken, E-Commerce und andere) zusammenarbeiten, um spezifische Standards und Richtlinien zu entwickeln, um sicherzustellen, dass sie in der Lage sind, auf die Herausforderungen der automatisierten Entscheidungsfindung durch Algorithmen zu reagieren und die Interessen der Verbraucher und der Öffentlichkeit zu berücksichtigen.“*

Der Bericht schließt ab mit Empfehlung 9:

> *„Der Europarat als führende Menschenrechtsorganisation des Kontinents ist der geeignete Ort, um die Auswirkungen der zunehmenden Nutzung automatisierter Datenverarbeitungs- und -entscheidungssysteme (insbesondere Algorithmen) auf öffentliche und private Bereiche auf die wirksame Ausübung der Menschenrechte weiter zu untersuchen. Sie sollte ihre Bemühungen in dieser Hinsicht fortsetzen, um geeignete Normensetzungsinstrumente für die Leitlinien für die Mitgliedstaaten zu entwickeln.“*

Die Empfehlungen 5 und 6 sind mit dem in unserem Bericht vorgeschlagenen Lösungen konsistent. Empfehlung 9, die dem Europarat besondere Kompetenzen zuweist, konfligiert potenziell mit Empfehlung 6, die Subsidiarität betont. Sollte der Rat Empfehlung 9 folgen und ein eigenes Zertifizierungssystem entwickeln, könnte es organisatorisch und inhaltlich zu Konflikten/Überschneidungen mit unseren Vorschlägen kommen. Bislang hat der Rat keine

[261] https://rm.coe.int/algorithms-and-human-rights-en-rev/16807956b5/.

Stellungnahme zu dem Bericht abgegeben, der erst im Juni vorgelegt wurde. Es muss auch bemerkt werden, dass die Analyse des Berichts auf einer sehr hohen Abstraktionsstufe blieb, mit wenig technischen Details, was auch die Zusammensetzung des Gremiums reflektiert. Wir halten es für unwahrscheinlich, dass in naher Zukunft durch den Europarat ein eigenständiges Zertifizierungssystem für Algorithmen entwickelt werden wird, das mit nationalstaatlichen Ansätzen in Konflikt geraten könnte.

Zusammenfassend lässt sich sagen, dass es trotz einer großen Anzahl von Initiativen eher unwahrscheinlich ist, dass Fragen der ADM-Kontrolle, -Analyse und -Zertifizierung auf europäischer Ebene vorweggenommen werden. Die Mehrzahl der Ansätze ist mit dem, was unsere Studie vorschlägt, konsistent und in der Tat komplementär. Dies ist besonders deutlich in den mehrfachen Forderungen nach neuer und zielgerichteter Forschung, einem Thema, das in allen diesen Studien zentral ist, wobei es aber generell an Details fehlt. Es scheint aber wahrscheinlich, dass es neue EU-geförderte Forschungsinitiativen geben wird, die insbesondere die technischen Aspekte der algorithmischen Transparenz betreffen werden. Insoweit auch unser Bericht Forschungsfragen aufwirft, die nicht nur nationales Recht betreffen, könnte hier ein Forschungsbiotop entstehen, das viele der auch von uns identifizierten Probleme ansprechen wird. Der zweite konsistent zu findende Vorschlag ist der von Zertifizierungen und Zertifizierungsanstalten. Auch hier fehlt es an Details, doch scheint es keine Pläne zu geben, mit nationalstaatlichen Ansätzen in Konkurrenz treten zu wollen.

7 Möglichkeiten der rechtlichen Regelung von Algorithmen in Deutschland

7.1 Herausforderungen und Möglichkeiten der rechtlichen Regelung fehlerhafter algorithmischer Entscheidungen

Gegenstand der Studie ist, wie eingangs dargestellt, die Frage, wie fehlerhafte Beurteilungen von Menschen durch Maschinen mit Hilfe rechtlicher Maßnahmen bekämpft werden können.

Die Optionen und Herausforderungen werden nachfolgend erörtert. Zunächst werden die Möglichkeiten der rechtlichen Regulierung, die derzeit intensiv diskutiert werden, anhand der aktuellen Diskussion zur Algorithmenregulierung dargestellt (7.1.1). Sodann werden, am Beispiel der Diskriminierung, zentrale Herausforderungen der rechtlichen Regelung identifiziert (7.1.2), bevor der rechtliche Rahmen von Tests (7.2) und von Transparenz (7.3) als wesentlichen Mitteln zur Bekämpfung problematischer algorithmischer Entscheidungen untersucht wird.

7.1.1 Algorithmenregulierung in der aktuellen Diskussion

Die Möglichkeiten zur gesetzlichen Regulierung von Algorithmen in Bezug auf deren spezifische Herausforderungen werden in Deutschland erst in jüngster Zeit intensiv diskutiert; die meisten Veröffentlichungen und Stellungnahmen stammen aus den Jahren 2017 und 2018.

Die Ansätze sind sehr unterschiedlich. So formuliert beispielsweise Wischmeyer in seiner umfangreichen Untersuchung „regulatorische Leitlinien“ für den Einsatz intelligenter Systeme und nennt als derartige Leitlinien etwa die „Sichtbarmachung der regulierenden Wirkung von intelligenten Systemen“, ein „angemessenes Qualitätsniveau für intelligente Systeme“, die Abwehr von „Diskriminierung durch intelligente Systeme“, „Datenschutz- und Informationssicherheit beim Einsatz intelligenter Systeme“, die „problemadäquate Verwendung von intelligenten Systemen“ sowie die „Haftungs- und Verantwortungsklarheit beim Einsatz intelligenter Systeme“.[262]

Eine breite Übersicht an Regelungsoptionen wurde in Deutschland in einem Forschungsprojekt der Universität Speyer unter der Leitung von *Martini* erstellt, dessen Ergebnisse noch nicht veröffentlicht sind. In einem vorab publizierten Beitrag von Martini[263] findet sich eine Zusammenfassung. In diesem Beitrag unterbreitet Martini eine Reihe von Regulierungsvorschlägen. Die Darstellung ist gegliedert in „Präventive Regulierungsinstrumente“, „Begleitende Fehlerkontrolle und Risikomanagement“, „Selbstregulierung“ und „Ex-post-Schutz“ und schlägt zahlreiche Maßnahmen vor.

[262] Wischmeyer, AöR 2018, 1, 18 ff.
[263] Martini, JZ 2017, 1017 ff.

Die Darstellung von Martini ist wegen der Vielzahl der vorgeschlagenen Instrumente für die Zwecke dieser Studie von besonderem Interesse. Daher werden nachfolgend, der Ordnung von Martini folgend, die in seiner Studie vorgeschlagenen Regelungsinstrumente genannt.

7.1.1.1 Anspruch auf eine Entscheidung durch natürliche Personen

Martini weist zustimmend auf den Grundgedanken des Art. 22 DSGVO hin, wonach Menschen nicht „zu reinen Objekten der Entscheidung einer Softwareanwendung herabgewürdigt“ werden dürften. Zugleich macht er darauf aufmerksam, dass der Anwendungsbereich des Art. 22 DSGVO wesentlich enger sei, als es die Überschrift vermuten lasse.[264] Diesen Überlegungen lässt sich entnehmen, dass ein weitergehender Anspruch auf eine Entscheidung durch eine natürliche Person geregelt werden könnte. Zu klären wäre dann auch, ob sich der Anspruch auch auf die Entscheidungsvorbereitung beziehen sollte. Dazu äußert sich Martini jedoch nicht.

In Bezug auf den Anspruch aus Art. 22 DSGVO schlägt Ernst die Einführung einer Protokollierungspflicht vor. Damit ist wohl die Pflicht gemeint, die Befassung eines menschlichen Entscheidungsträgers zu protokollieren.[265]

7.1.1.2 Kennzeichnungspflicht für algorithmenbasierte Verfahren

Wirksamer Rechtsschutz gegen algorithmenbasierte Entscheidungen setzt nach Martini die Kenntnis des Betroffenen voraus, von einer solchen Entscheidung tangiert zu sein. Daher schlägt Martini vor, eine gesetzliche Kennzeichnungspflicht, die in Ansätzen in Art. 13 Abs. 2 lit. f) und Art. 41 Abs. 2 lit g) DSGVO existiere, „auf alle algorithmenbasierte Verfahren in persönlichkeitssensiblen Feldern zu erstrecken“. Die Kennzeichnung sollte „mithilfe visuell leicht erfassbarer Symbole erfolgen“[266].

7.1.1.3 Begründungspflicht für algorithmenbasierte Entscheidungsverfahren

Martini erwägt eine Begründungspflicht für algorithmenbasierte Entscheidungsverfahren, die deren Intransparenz entgegenwirken könne. Dieser Vorschlag entspricht teilweise der Debatte zu einem „Recht auf Erklärung“,[267] dessen Leistungsfähigkeit in der juristischen Literatur teilweise aber auch kritisch gesehen wird.[268]

Da eine umfassende Begründungspflicht Programmierer vor erhebliche Herausforderungen stelle, schlägt Martini eine differenzierende Regelung vor, bei der die gesetzlich geforderte Begründungstiefe mit dem Risiko von Diskriminierung und Persönlichkeitsgefährdung korreliert. Eine allgemeine Pflicht zur Offenlegung des Programmcodes für Software hingegen lehnt Martini ausdrücklich ab.[269]

[264] Martini, JZ 2017, 1017, 1019 f.
[265] Ernst, JZ 2017, 1026, 1031.
[266] Martini, JZ 2017, 1017, 1020.
[267] Siehe dazu Wischmeyer, AöR 2018, 1, 48 ff. mit weiteren Nachweisen.
[268] Vgl. etwa die skeptische Einschätzung von Wischmeyer, AöR 2018, 1, 52 ff.
[269] Martini, JZ 2017, 1017, 1020.

7.1.1.4 Transparenzanforderungen an algorithmenbasierte Dienste der Nachrichtenauswahl

In Bezug auf Nachrichtenaggregatoren, für die beispielhaft Google News und Newsfeed von Facebook genannt werden, schlägt Martini weitergehende Transparenzpflichten dahingehend vor, dass die Anbieter von Diensten oberhalb einer „kritischen meinungsbildungsrelevanten Größenschwelle“ verpflichtet seien, „einen öffentlichen Einblick in ihr technisches Verfahren der Nachrichtenauswahl“ zu geben.[270]

7.1.1.5 Inhaltliche Ex-ante-Kontrolle bei algorithmenbasierten Entscheidungsverfahren

Ein offensichtliches Mittel zur Regulierung algorithmenbasierter Entscheidungsverfahren ist eine staatliche Vorabkontrolle von Software. Martini votiert insoweit dafür, für behördliche Softwareanwendungen eine Vorabprüfung zwingend vorzusehen.[271] Für sonstige Anwendungsbereiche sollte dies dann sein, wenn Software in „besonders persönlichkeitssensiblen Bereichen“ eingesetzt werden soll. Gegenstand einer solchen Kontrolle sollen der „Programmcode deterministischer Verfahren“, „die korrekte Einbindung der Datenbasis“ sowie bei lernfähiger Software auch der Trainingsprozess sein.[272]

7.1.1.6 Allgemeines Gleichbehandlungsgesetz (AGG)

Das AGG, das vor Diskriminierung schützen soll, ist technologieneutral konzipiert und daher auf softwarebasierte Verfahren anwendbar. Martini weist insoweit darauf hin, dass der Anwendungsbereich des AGG auf bestimmte Lebensbereiche beschränkt ist und insbesondere bei Verträgen zwischen Privaten zahlreiche Anwendungsfälle softwarebasierter Verfahren nicht umfasst. Als Ausweg empfiehlt Martini die Ergänzung des § 2 Abs. 2 AGG um eine neue Nr. 9 für „Ungleichbehandlungen zwischen Privaten […], die auf einer algorithmenbasierten Datenauswertung oder einem automatisierten Entscheidungsverfahren beruhen.“[273]

7.1.1.7 Kontrollalgorithmen und Kontrollpflichten

Software unterliegt typischerweise einem dynamischen Veränderungsprozess, womit jede punktuelle Kontrolle systematisch geschwächt wird. Rechtskräftige Urteile erreichen, worauf Martini hinweist, häufig einen nicht mehr verwendeten Softwarestand.[274]

Wegen der dynamischen Veränderungsprozesse von Software schlägt Martini daher die Einführung von „Kontrollpflichten“ vor. Dabei ist offenbar an eine aufsichtsrechtliche Prüfung gedacht, die sich bei lernfähigen Systemen auf Trainingsumgebung, Testdaten und Richtigkeit der Datenbasis erstrecken soll.

Als Prüfwerkzeug verweist Martini auf „Kontrollalgorithmen“, die die Entscheidungsergebnisse der Software analysieren. Zum Schutz berechtigter Geheimhaltungsinteressen könnten vertraulichkeitswahrende Instrumente, insbesondere In-

[270] Martini, JZ 2017, 1017, 1021.
[271] Ebenso Martini/Nink, NVwZ-Extra 20/2017, 1, 12.
[272] Martini, JZ 2017, 1017, 2021.
[273] Martini, JZ 2017, 1017, 1021.
[274] Martini, JZ 2017, 1017, 2021.

camera-Verfahren, vorgesehen werden.[275] Ähnlich wird in der Literatur für den Bereich der automatisierten Verwaltungsverfahren eine Pflicht zur fortlaufenden Überprüfung von Systemen Maschinellen Lernens als notwendig betrachtet.[276]

7.1.1.8 Risikomanagementsysteme und Veröffentlichung der Risikoabschätzung

In „persönlichkeitssensiblen Einsatzbereichen automatischer Entscheidungsmechanismen" empfiehlt Martini die gesetzliche Verpflichtung zur Einführung von Risikomanagementsystemen. Als Beispiel für solche Einsatzbereiche nennt Martini insbesondere automatische Verwaltungsverfahren. Dabei mahnt Martini regulatorisches Augenmaß an und ruft den Gesetzgeber auf, einen Ausgleich zwischen wirtschaftlichen und persönlichkeitsrechtlichen Interessen herzustellen.[277]

7.1.1.9 Protokollierung der Programmabläufe

Um die Möglichkeit zu gewährleisten, etwaige Rechtsverstöße zu überprüfen, befürwortet Martini eine „umfangreiche Protokollierung der Programmabläufe und Rückkopplungsprozesse", die wohl durch Gesetz verpflichtend gemacht werden soll.[278] Im Hinblick auf die etwaige Überforderung von Softwarenutzern solle die Protokollierungspflicht aber hinsichtlich ihres Anwendungsbereich und Umfangs der Protokollierungspflicht nach dem Schutzbedarf und der „Skalierungsintensität des Geschäftsmodells" differenzieren.[279]

7.1.1.10 Selbstverpflichtung nach dem Vorbild des Corporate Governance Kodex

Um den Sachverstand der Anbieter von Software einzubeziehen, schlägt Martini die Nutzung staatlich kontrollierter Selbstregulierungsinstrumente nach dem Vorbild des Corporate-Governance-Kodex und dessen Einbettung in § 161 AktG vor. Insofern sollen sich Anbieter „besonders persönlichkeitssensibler, insbesondere lernfähiger Software" in einem „Algorithmic Responsibility Kodex" erklären müssen.[280]

7.1.1.11 Beweislastregelung für den Zivilprozess

Im Hinblick auf die Nachweisschwierigkeiten für Betroffene im Zivilprozess empfiehlt Martini die Einführung eines abgestuften Systems der Beweislastverteilung. Danach solle es zugunsten des Betroffenen genügen, wenn er Tatsachen vortrage, die mit „überwiegender Wahrscheinlichkeit" auf eine Diskriminierung schließen ließen. Der Anbieter müsse dann den Gegenbeweis des Nichtvorliegens einer Rechtsverletzung führen.[281]

7.1.1.12 Gefährdungshaftung mit Versicherungspflicht

Für Softwareanwendungen mit besonderen Risiken erwägt Martini die Einführung einer Gefährdungshaftung. Als Beispiel für derartige sensible Anwendungen nennt er „medizinische Anwendungen" oder „Pflegeroboter".[282]

[275] Martini, JZ 2017, 1017, 2022.
[276] So etwa Martini/Nink, NVwZ-Extra 20/2017, 1, 12.
[277] Martini, JZ 2017, 1017, 2022.
[278] Ebenso, für automatisierte Verwaltungsverfahren, Martini/Nink, NVwZ-Extra 20/2017, 1, 12.
[279] Martini, JZ 2017, 1017, 1022.
[280] Martini, JZ 2017, 1017, 1022 f.
[281] Martini, JZ 2017, 1017, 1024.
[282] Martini, JZ 2017, 1017, 1024.

7.1.1.13 Erweiterung von Abmahnbefugnissen nach dem Gesetz gegen den unlauteren Wettbewerb (UWG)

Um die Expertise von Wettbewerbern und deren Interesse an der Unterbindung unzulässiger Geschäftspraktiken zu nutzen, schlägt Martini die Erweiterung des UWG um den Tatbestand „diskriminierender oder sonst persönlichkeitsverletzender Softwareanwendungen“ vor.[283]

7.1.1.14 Verbandsklagerecht der Verbraucherverbände und Schiedsstelle

Als Instrument vor allem zum Schutz von Verbraucherinteressen erwägt Martini die Schaffung eines Verbandsklagerechts der Verbraucherschutzverbände nach dem Unterlassungsklagegesetz (UKlaG) für Softwareanwendungen in „persönlichkeitssensiblen Anwendungsfeldern“. Verbraucherverbände könnten dann gegen „rechtswidrige, insbesondere diskriminierende algorithmenbasierte Entscheidungsfindung vorgehen“. Als weiteres Instrument des Verbraucherschutzes nennt Martini eine staatlich geförderte Schlichtungsstelle.[284]

7.1.1.15 Nebenfolgenkompetenz der Zivilgerichte

Als Schutzinstrument erwägt Martini schließlich die gesetzliche Erweiterung der Wirkung zivilgerichtlicher Entscheidungen, etwa die Ausdehnung der Rechtskraft auf Dritte.[285]

7.1.1.16 Zwischenergebnis

Die in der Untersuchung von Martini genannte Vielzahl möglicher Regulierungsoptionen belegt, dass das Problem durch ganz unterschiedliche rechtliche Mittel und Konzepte angegangen werden kann. Die Regulierungsoptionen, deren Aufzählung in der Studie von Martini im Wesentlichen vollständig ist, umfasst alle in Betracht kommenden Handlungsformen des Gesetzgebers.

Eindrucksvoll ist, dass Martini auch den Einsatz aller genannten Instrumente empfiehlt, und zwar offenbar kumulativ. Dies führt im Ergebnis zum Vorschlag einer umfassenden Regulierung von Algorithmen auf allen Ebenen. Nach diesem Ansatz wäre ein spezifischer Rechtsrahmen für ADM-Systeme oder „Algorithmen“ im Allgemeinen unter Einschluss von Aufsicht, Haftungsrecht und prozessualen Sonderregeln zu schaffen. Andere Stellungnahmen hingegen favorisieren eine eher punktuelle Regelung oder warnen gar vor einer Überregulierung der neuen Technologie.

Es ist offensichtlich, dass der derzeitige Stand der Forschung eine endgültige Festlegung auf einen spezifischen Rechtsrahmen für ADM-Systeme oder „Algorithmen“ nicht erlaubt. Vielmehr ist es erforderlich, die Leistungsfähigkeit der einzelnen rechtlichen Instrumente im Lichte der spezifischen Herausforderungen der neuen Technologie näher zu erforschen.

Angesichts dieses Diskussions- und Forschungsstandes soll diese Studie weder einen erneuten Überblick über Regelungsoptionen geben noch versuchen, einen umfassenden rechtlichen Rahmen zu beschreiben. Vielmehr sollen Lösungsansätze für zentrale rechtliche Herausforderungen von ADM-Systemen am Beispiel der Diskriminierung herausgearbeitet werden.

[283] Martini, JZ 2017, 1017, 1024.
[284] Martini, JZ 2017, 1017, 1024 f.
[285] Martini, JZ 2017, 1017, 1025.

7.1.2 Herausforderungen der Regelung von ADM-Systemen am Beispiel der Diskriminierung

7.1.3 Herausforderungen der rechtlichen Regelung von Diskriminierung

Eine wesentliche Herausforderung der Regelung der Diskriminierung durch ADM-Systeme ergibt sich dadurch, dass bereits die rechtliche Regelung von Diskriminierung im Allgemeinen überaus komplex ist und zahlreiche ungeklärte Rechtsfragen aufweist. Als Ergebnis der Untersuchung (5.4) sind für die rechtliche Regelung einer Diskriminierung durch Algorithmen neben der Schwierigkeit, eine Diskriminierung zu erkennen, auch folgende Aspekte von Bedeutung, was darauf beruhen dürfte, dass eine rechtliche Regelung der Diskriminierung großen Herausforderungen begegnet.

7.1.3.1 Abgrenzung von Diskriminierung und zulässiger Ungleichbehandlung

So ist die Abgrenzung einer rechtlich unzulässigen Diskriminierung von einer rechtlich zulässigen oder gar gebotenen Ungleichbehandlung sehr schwierig. Ungleichbehandlung ist nicht per se unzulässig, vielfach sogar gefordert, und auch Ungleichbehandlung aufgrund von Merkmalen, wie dem Geschlecht, die in manchem Zusammenhang als unzulässige Diskriminierung zu untersagen ist, ist in einem anderen Zusammenhang erlaubt.

Soweit eine Diskriminierung oder Ungleichbehandlung durch gesetzliche Instrumente bekämpft werden soll, stellen sich dieselben Abgrenzungsfragen in Bezug auf jedes einzelne Instrument, also etwa die Frage, in welchen Bereichen eine Behörde eingreifen soll oder ein Verbraucherschutzverband Klagebefugnisse erhalten soll.

7.1.3.2 Einbettung der Diskriminierung im Rechtssystem

Die Komplexität der rechtlichen Erfassung von Diskriminierung ergibt sich vor allem daraus, dass Diskriminierung Gegenstand ganz unterschiedlicher Rechtsbereiche, namentlich des Zivilrechts und des öffentlichen Rechts, ist und sehr unterschiedlich eingebettet sein kann.

Sowohl im Bereich des Tatbestands einer Diskriminierung als auch in Bezug auf Rechtsfolgen bestehen ganz unterschiedliche Ansätze. Vereinzelt bestehen Sonderregeln speziell zur Diskriminierung, etwa in Form des AGG, das Diskriminierung in einem recht begrenzten Ausschnitt vertraglicher Beziehungen erfasst. Das Verbot der Diskriminierung ergibt sich daneben aber auch aus allgemeinen Regeln. Beispielhaft sei die deliktische Generalklausel des § 823 Abs. 1 BGB genannt, die nicht ausdrücklich auf Diskriminierung Bezug nimmt, aber bei Diskriminierung unter dem Gesichtspunkt der Persönlichkeitsrechtsverletzung einen Schadensersatzanspruch gewährt.[286]

Vielfach ist Diskriminierung als Sonderfall einer unzulässigen Entscheidung zu betrachten und wird damit durch allgemeine Anforderungen an Entscheidungen erfasst. So ist beispielsweise die Personalauswahl im Beamtenrecht umfassend reguliert. Die diskriminierende Auswahlentscheidung ist nur ein Sonderfall der fehlerhaften Personalauswahl und wird durch die allgemeinen Regeln erfasst.

[286] OLG Köln, NJW 2010, 1676, 1677; Dauner-Lieb/Langen-Katzenmeier, § 823 BGB Rn. 237; Staudinger-Olzen, Einleitung zum Schuldrecht, Rn. 234.

Zur Durchsetzung des Verbots werden *de lege lata* unterschiedliche Instrumente eingesetzt: Zivilrechtlich bestehen Schadensersatz- und Unterlassungsansprüche. So gewährt die deliktische Generalklausel des § 823 Abs. 1 BGB einen Schadensersatzanspruch. Im öffentlichen Bereich gelten ergänzend Regeln der Beamtenhaftung und des Staatshaftungsrechts.

Neben dem Anspruch auf Schadensersatz können Ansprüche auf Beseitigung und Unterlassung bestehen, etwa aus § 1004 BGB (oder den §§ 823 Abs. 1, 1004 BGB) analog unter dem Gesichtspunkt der Persönlichkeitsrechtsverletzung. Im Rahmen rechtlicher Sonderbeziehungen liegt bei (unzulässiger) Diskriminierung regelmäßig auch die Verletzung einer vertraglichen Nebenpflicht vor, die nach § 280 Abs. 1 BGB zu einem Anspruch auf Schadensersatz führen kann.

Eine spezifische öffentlich-rechtliche Kontrolle der Diskriminierung, etwa durch eine spezielle Aufsicht gegen Diskriminierung, besteht nach geltendem Recht nicht. Vielmehr kann Diskriminierung, soweit überhaupt, ausschließlich im Rahmen sonstiger Aufsicht erfasst werden.

7.1.4 Kernprobleme der Diskriminierung durch Algorithmen

Diese – enormen – Herausforderungen im Umgang des Rechts mit Diskriminierung können als solche nicht Gegenstand dieser Studie sein, auch wenn die Diskussion und Forschung im Zusammenhang mit Diskriminierung durch Algorithmen von diesen Schwierigkeiten in hohem Maße betroffen ist. So stoßen viele technische Überlegungen aufgrund der Unklarheit und Komplexität der Rechtslage der Diskriminierung an ihre Grenzen. Hieraus ergeben sich auch de lege ferenda besondere Herausforderungen, wie sich an der Untersuchung der einzelnen rechtlichen Mittel zur Bekämpfung von Diskriminierung deutlich zeigen wird.

Für die Zwecke der Studie ist es notwendig, die Überlegungen auf den Kern der algorithmenspezifischen Problematik zu fokussieren. Daher werden in Bezug auf die aufgezeigten Schwierigkeiten durch vereinfachende Annahmen Beschränkungen getroffen.

Die schwierige Abgrenzung von Diskriminierung und gerechtfertigter Ungleichbehandlung ist als solche keine algorithmenspezifische Fragestellung und wird daher aus der weiteren Untersuchung (siehe unten Kapitel 7.2 ff.) ausgeklammert. Insoweit soll als Vereinfachung zugrunde gelegt werden, dass es im geltenden Recht einen Bestand an rechtlich unzulässiger Diskriminierung gibt.

Ebenso ist die Frage, in welchen Bereichen ein spezifisches rechtliches Instrument der Diskriminierungsbekämpfung, etwa ein Gleichbehandlungsgebot nach Art des AGG, einzusetzen ist, kein Spezifikum der Diskriminierung durch Algorithmen und soll im Rahmen dieser Studie außer Betracht bleiben, auch wenn sich hierdurch erhebliche Limitierungen in Bezug auf die Konkretisierung möglicher Handlungsempfehlungen ergeben.

Die Einbettung der Diskriminierung in das Rechtssystem kann aufgrund ihrer ungeheuren Komplexität in dem begrenzten Rahmen dieser Studie in keiner Weise vollständig erfasst werden. Sie kann aber auch nicht völlig außer Acht gelassen werden, da die Verschiedenheit der rechtlichen Mechanismen gerade auch bei Algorithmen von entscheidender Bedeutung ist. Die Studie bezieht daher strukturell unterschiedliche Mittel, insbesondere zivilrechtliche Pflichten sowie Beweisregeln, am Beispiel der Diskriminierung in die Betrachtung ein.

Gegenstand der Studie sollen Kernprobleme der Diskriminierung durch Algorithmen sein – hier, wie eingangs dargestellt, die Frage, wie fehlerhafte Beurteilungen von Menschen durch Maschinen mit Hilfe rechtlicher Maßnahmen bekämpft werden können.

Ein wesentliches Problem wird hier in der Feststellung einer Diskriminierung bei maschineller Beurteilung von Personen gesehen. Dieses Problem besteht schon bei herkömmlichen Systemen, die einem vollständig vorprogrammierten Schema folgen, vor allem aber bei Systemen, die mit Verfahren des *Machine Learning* arbeiten. Diese Problematik soll daher unter besonderer Berücksichtigung der Probleme bei *Machine Learning* analysiert werden.

7.2 Feststellung fehlerhafter Beurteilung am Beispiel der Diskriminierung

7.2.1 Feststellung von Diskriminierung bei Beurteilungen

Eine wesentliche Herausforderung jeder rechtlichen Regelung fehlerhafter Beurteilung von Personen ist die Feststellung eines Fehlers. Im Fall der Diskriminierung beispielsweise ist die Feststellung erforderlich, dass ein unzulässiges Merkmal einen Einfluss auf die Entscheidung hatte.

7.2.1.1 Determinierte und undeterminierte Entscheidungen

Diese Schwierigkeit besteht insbesondere bei Entscheidungen, die von Menschen getroffen werden, da der eigentliche Entscheidungsweg oft nicht nachvollziehbar und meist selbst dem Entscheider nicht in vollem Umfang bewusst ist.[287] Soweit etwa „Erfahrung“ oder „Intuition“ eines Entscheiders von Bedeutung sind, erfolgen wesentliche Entscheidungsschritte unbewusst. Sie sind damit einer unmittelbaren Analyse nicht zugänglich. Auch bei von Menschen getroffen Entscheidungen kann es freilich so liegen, dass der Prozess der Entscheidungsfindung eindeutig und transparent ist, etwa wenn der Entscheider einem vorgegebenen und schriftlich fixierten Schema folgt. So ist bei Auswahlentscheidungen häufig vorgegeben, welche Kriterien mit welchem Gewicht zu berücksichtigen sind (z.B. 30 % Examensnote, 20 % Wartezeit, nach Semestern). Auch wenn hier ein Mensch eine Entscheidung trifft, so folgt diese streng vorgegebenen Kriterien und ist vollständig transparent.

Es ist also auch bei menschlichen Beurteilungen von Personen zwischen zwei Grundtypen von Entscheidungen zu differenzieren, die hier bewusst außerhalb sonstiger fachlicher Terminologien als „determiniert“ und „undeterminiert“ bezeichnet werden sollen. Zum einen gibt es die vollständig determinierte Entscheidung, deren Parameter und Gewicht der Parameter in vollem Umfang vorgegeben sind. Zum anderen gibt es Entscheidungen, deren Entscheidungsweg nicht vollständig vorgegeben ist und bei denen der Entscheider auf Kriterien und Gewichtungen zurückgreift, die weder vorbestimmt noch vollständig bekannt sind und unter Umständen auch dem Entscheider nicht bekannt (bewusst) sind. Sämtliche komplexen Beurteilungsentscheidungen gehören zur zweiten Gruppe.

[287] Darauf weisen etwa auch Martini/Nink, NVwZ-Extra 10/2017, 1, 10 für diskriminierende Entscheidungen hin.

Die beiden gedanklich scharf zu trennenden Entscheidungsarten werden in der Praxis sehr häufig kombiniert und gehen letztlich sogar ineinander über, soweit im Rahmen vordefinierter Kriterien wiederum Bewertungen erforderlich sind. Dessen ungeachtet handelt es sich um ganz unterschiedliche Arten von Entscheidungen, die rechtlich auch mit sehr unterschiedlichen Mitteln angesprochen werden.

Bei algorithmischen Entscheidungen, also Beurteilungen durch Maschinen, ist letztlich dieselbe Unterscheidung zu treffen. Entsprechend ist zu fragen, ob eine vollständig vorgegebene Entscheidung vorliegt oder ob eine Maschine eine Entscheidung trifft, die vom Nutzer nicht vollständig vorhersehbar oder vorgegeben ist. Dies ist insbesondere der Fall, wenn sie nicht oder nicht mit vertretbarem Aufwand durch Einsicht in einen Programmcode überprüft werden kann.

7.2.1.1.1 Qualitätssicherung und Überprüfung undeterminierter Entscheidungen

Die undeterminierte Entscheidung ist auch aus rechtlicher Sicht eine zentrale Herausforderung, die mit verschiedenen Mitteln angegangen wird. Zum einen kennt das Recht ein umfangreiches Arsenal von Möglichkeiten, um die Qualität solcher undeterminierten Entscheidung zu sichern.

So werden häufig Anforderungen an die Qualität des Entscheiders geregelt, etwa in Form einer bestimmten Ausbildung des Entscheiders (z.B. Anforderungen an Richter), sowie an dessen Entscheidungssituation (Gebot der Unabhängigkeit, Ausschluss bei Befangenheit etc.). Bei wichtigen Entscheidungen wird zur Qualitätssicherung gar eine Kollegialentscheidung angeordnet.

Auf die vollständige Klärung von Entscheidungsfehlern wird häufig verzichtet. Vielmehr wird das Risiko durch Beweisregeln verteilt. So genügt mitunter der Verdacht einer fehlerhaften Entscheidung (z.B. einer Diskriminierung), um Rechtsfolgen oder jedenfalls eine Beweislastumkehr auszulösen.

Auf der anderen Seite gewährt das Recht in dieser Situation regelmäßig einen sogenannten Beurteilungsspielraum, der der Überprüfung undeterminierter Entscheidungen Grenzen setzt.

Die rechtliche Regelung von Diskriminierung reagiert hierauf durch ganz unterschiedliche Instrumente, nicht zuletzt durch Beweisregelungen im weitesten Sinne, bis hin zu sehr scharfen Eingriffen wie Vermutungen mit Beweislastumkehr. Aber auch Dokumentationspflichten gehören zum Arsenal des Rechts, ebenso Regeln zur Datenerhebung. So ist es in vielen Zusammenhängen unzulässig, bestimmte Informationen über eine Person zu erheben. Das vielleicht bekannteste Beispiel ist die unzulässige Frage nach Schwangerschaft oder Kinderwunsch bei einem Einstellungsgespräch.

Bei maschinellen Beurteilungen bestehen im Ausgangspunkt dieselben Schwierigkeiten wie bei Beurteilungen, die von Menschen vorgenommen werden. Soweit undeterminierte Entscheidungen von Maschinen gefällt werden, ist auch hier zu erwägen, einen Beurteilungsspielraum zuzubilligen mit der Folge, dass die Entscheidung nicht als fehlerhaft gilt, wenn sie innerhalb dessen bleibt.

7.2.1.2 Zwischenergebnis und Fragestellung der Studie

Die Abgrenzung dieser Fallgruppen ist nicht Aufgabe dieser Studie. Zur Vereinfachung wird daher angenommen, dass es einerseits Fälle gibt, in denen eine Beurteilungsentscheidung

durch Einsicht in den Programmablauf mit verhältnismäßigem Aufwand nachvollzogen werden kann, und andererseits Fälle, in denen dies nicht der Fall ist.

Aus den vorgenannten Überlegungen ergibt sich als Fragestellung für diese Studie, ob und mit welchen Mitteln bei determinierten Entscheidungen einerseits, undeterminierten (Beurteilungs-)Entscheidungen andererseits ein Fehler der Beurteilung festgestellt werden kann. Diese Frage wird hier am Beispiel der Diskriminierung untersucht.

7.2.2 Mittel der Feststellung von Diskriminierung durch maschinelle Beurteilungen

Zur Feststellung von Diskriminierung durch maschinelle Beurteilungen werden vor allem zwei Mittel diskutiert: die Analyse des Programms, dem die Entscheidung folgt, und das Testen des Systems auf Diskriminierung.

7.2.2.1 Code-Analyse

Als geeignetes Mittel zur Feststellung von Diskriminierung durch Algorithmen wird in der Literatur, nicht zuletzt in der Studie von Martini, die Code-Analyse genannt. Damit ist gemeint, dass durch Einsicht in das Computerprogramm, das die Beurteilung der Person vornimmt, untersucht wird, ob eine Diskriminierung im Sinne einer unmittelbaren Benachteiligung aufgrund eines unzulässigen Merkmals vorliegt.

Die Code-Analyse zur Ermittlung einer etwaigen Diskriminierung steht vor erheblichen Herausforderungen:

- Eine solche Untersuchung kann sehr anspruchsvoll sein: Sie kann sehr aufwendig sein, insbesondere wenn das Computerprogramm umfangreich ist und die Verwendung des Merkmals nicht offensichtlich ist. In jedem Fall setzt eine solche Analyse Fachkenntnisse voraus und wird daher regelmäßig nur von Experten durchgeführt werden können.

- Bei Verfahren, die Technologien des *Machine Learning* nutzen, läuft die Code-Analyse weitgehend ins Leere. Wie bereits in Kapitel 3.1 erklärt, müssen solche Modelle in vielen Fällen als intransparente Blackboxes angesehen werden, die lediglich auf Testdaten statistisch evaluiert werden können. Dies wiederum setzt Testdaten voraus, die entsprechend bestimmten Kriterien der technischen Kompatibilität, statistischen Repräsentativität, Aktualität und Tauglichkeit als Prüfungsmaßstab genügen müssen. Falls trainierte Modelle als einsehbare Whitebox zur Verfügung stehen, ist eine Expertenprüfung zwar prinzipiell möglich, jedoch mit großem Aufwand verbunden und zum nächsten Modelltrainingszyklus im Regelfall obsolet. Die Möglichkeit, dies durch strukturierte Prozesse praktikabel zu gestalten, wird im Gutachten entsprechend untersucht (siehe Kapitel 4.4).

- Ein wesentliches Hindernis der Feststellung von Diskriminierung durch Code-Analyse ist die Verfügbarkeit des Computerprogramms zur Einsichtnahme und Analyse. Die von der Beurteilung betroffene Person hat typischerweise selbst keinen Zugriff auf das Programm, das meist im Besitz des Vertragspartners oder eines Dritten, etwa eines Dienstleisters, steht. Es bedarf mithin eines Rechts auf Einsicht in oder Herausgabe des Programms zur Untersuchung. Ein solches Recht besteht derzeit weder in Form eines zivilrechtlichen Anspruchs auf Einsichtnahme oder Herausgabe des Programms noch in Form einer expliziten Einsichtsbefugnis einer Behörde. Als

maßgeblicher Stand des Rechts kann wohl die Entscheidung des BGH zum Schufa-Scoring angesehen werden, in der der BGH die Frage, ob sich aus § 34 Abs. 4 BDSG a.F. ein Anspruch auf Kenntnis des Scoringprogramms der Schufa ergibt, mit Blick auf den Schutz des Programms als Geschäftsgeheimnis verneinte.[288]

- Der BGH stützt seine Entscheidung auf den Willen des Gesetzgebers, Geschäftsgeheimnisse der Auskunfteien zu schützen. Den Umfang des vom Auskunftsanspruch nach § 34 BDSG a. F. ausgenommenen Geschäftsgeheimnisses beschreibt der BGH wie folgt: „Zu den nach dem gesetzgeberischen Willen als Geschäftsgeheimnis geschützten Inhalten der Scoreformel zählen damit die im ersten Schritt in die Scoreformel eingeflossenen allgemeinen Rechengrößen, wie etwa die herangezogenen statistischen Werte, die Gewichtung einzelner Berechnungselemente bei der Ermittlung des Wahrscheinlichkeitswerts und die Bildung etwaiger Vergleichsgruppen als Grundlage der Scorekarten.“. Die h.M. der Literatur teilt die Auffassung des BGH.[289]

- Als Stand des geltenden Rechts wird man daher annehmen müssen, dass zivilrechtliche Auskunftsansprüche als solche regelmäßig nicht bestehen. Dies wirkt sich auch auf die Untersuchung im Rahmen zivilrechtlicher Verfahren aus. Im Zivilprozess gilt entsprechend dem Beibringungsgrundsatz, wonach die Parteien den Streitstoff vollständig vortragen müssen, dass die beweisbelastete Partei die Diskriminierung durch die maschinelle Beurteilung substantiiert vortragen muss. Soweit dazu eine Code-Analyse genutzt werden muss, muss diese also zur Vorbereitung des Prozesses vorliegen. Eine Auskunft vor Vorbereitung einer Klage oder zur Erstellung eines schlüssigen Vortrags kann nach deutschem Prozessrecht, anders als nach Prozessordnungen, die etwa eine „pre-trial discovery“ kennen, nicht verlangt werden.

Zusammengefasst ergeben sich für das Mittel der Code-Analyse folgende wesentliche Herausforderungen:

1.) Die Code-Analyse kann sehr aufwendig sein. Die Kosten können, bei Überprüfung in einem Einzelfall, in einem ungünstigen Verhältnis zum Gegenstandswert stehen.

2.) Die Code-Analyse läuft bei selbstlernenden Systemen weitgehend ins Leere, soweit nicht einfach nachvollziehbare Modelle (wie z.B. die logistische Regression) verwendet werden.

3.) Die Untersuchung des Codes zur Feststellung von Diskriminierung ist aus rechtlichen Gründen meist nicht möglich.

7.2.2.2 Tests

Als Mittel zur Feststellung von Diskriminierung wird in der aktuellen Forschung vor allem das Testen von Programmen genannt. Ein wesentlicher Vorteil dieses Konzepts ist, dass es auch bei selbstlernenden Systemen zur Anwendung kommt und damit eine wesentliche Schwäche der Code-Analyse vermeidet.

[288] BGH, Urteil vom 28.01.2014 – VI ZR 156/13 –, BGHZ 200, 38.

[289] Bräutigam/Schmidt-Wudy, CR 2015, 57, 62; Conrad, in: Auer-Reinsdorff/Conrad, § 34 Rn. 498; Dix, in: Simitis, § 34 Rn. 33; Gola/Klug/Körffer, in: Gola/Schomerus, BDSG, § 34 Rn. 12f; Kugelmann, DuD 2016, 566, 568; Martini, in: Paal/Pauly, Art. 22 Rn. 22; Paal/Hennemann, in: Paal/Pauly, Art. 13 Rn. 31; Taeger, MMR 2014, 492, 493.

Auch das Testen von Systemen auf Diskriminierung steht vor Herausforderungen. Geeignete standardisierte Testverfahren stehen, soweit ersichtlich, derzeit nicht in jedem Fall oder gar nicht zur Verfügung. In Bezug auf *Machine Learning* sind Testverfahren zurzeit Gegenstand der Forschung (siehe Kapitel 4.3.3).

Die Entwicklung und Durchführung eines Tests für eine spezielle Anwendung ist möglich, kann aber mit erheblichem Aufwand verbunden sein. Allerdings ist zu vermuten, dass bei traditionellen Programmen der Aufwand für die Feststellung von Diskriminierung anhand von Testdaten weniger hoch ist als der für die Code-Analyse.

Auch beim Testen von Systemen bestehen rechtliche Hindernisse. Die Durchführung eines Tests bedarf der Möglichkeit, das zu untersuchende System zu nutzen. Dies ist häufig nicht gegeben, wie das Beispiel des Schufa-Scorings demonstriert: Das Scoring-Programm wird von der Schufa für Testzwecke nicht zur Verfügung gestellt. Dies ist kein Ausnahmefall, sondern dürfte der praktische Regelfall sein.

Ein zivilrechtlicher Anspruch auf Nutzung eines Programms zu Testzwecken besteht *de lege lata* nicht.

7.2.2.3 Zwischenergebnis

Zur Feststellung von Diskriminierung bei maschineller Beurteilung von Menschen können insbesondere Code-Analysen und Tests eingesetzt werden. Beide Mittel haben aus technischer Sicht ihre Anwendungsbereiche, auch wenn Tests aus technischer Perspektive Vorteile haben, da sie vermutlich weniger aufwendig und zudem auch bei selbstlernenden Systemen anwendbar sind.

7.2.3 Feststellung von Diskriminierung de lege ferenda

Zur Lösung des Problems, dass die Feststellung von Diskriminierung bei maschinellen Beurteilungen von Menschen *de lege lata* schwierig ist, sind sehr unterschiedliche Wege denkbar, die im geltenden Recht teilweise bereits beschritten wurden. Nachfolgend werden, ohne Anspruch auf Vollständigkeit, zentrale Mittel untersucht.

7.2.3.1 Feststellung der Diskriminierung durch Indizien

Nach geltendem Recht wird bei maschinellen wie auch bei menschlichen Entscheidungen das Vorliegen einer Diskriminierung regelmäßig anhand von Indizien ermittelt. In Gerichtsverfahren wird meist auf den Indizienbeweis zurückgegriffen, der insoweit durchaus leistungsfähig ist.

Ein illustratives Beispiel ergibt sich aus einem Urteil des OLG Köln aus dem Jahre 2010 in einem Rechtsstreit zwischen einem Mietinteressenten und der Wohnungsverwaltungsgesellschaft.[290] Hier hatte die Hausmeisterin bei einer Wohnungsbesichtigung zum Zwecke des etwaigen Abschlusses eines Mietvertrags Folgendes gesagt: „Die Wohnung wird nicht an Neger, äh … Schwarzafrikaner und Türken vermietet“. Diese Aussage der Hausmeisterin war ein wesentliches Indiz für das Vorliegen einer Diskriminierung, auch wenn letztlich andere Gesichtspunkte von Bedeutung waren. Es wurde im Prozess durch die Vernehmung einer weiteren Zeugin aber unstreitig, dass die

[290] OLG Köln, Urteil v. 19.01.2009 – 24 U 51/09 –, NJW 2010, 1676.

Eigentümer der Wohnung eine Vermietung an farbige Mieter ablehnten.[291] Was bei Lektüre des Urteils deutlich wird, ist der Umstand, dass ohne diese Äußerung eine Diskriminierung nicht hätte festgestellt werden können, da nur aufgrund der Bemerkung der Hausmeisterin so substantiiert vorgetragen werden konnte, dass es zur Beweisaufnahme und damit zum Bekanntwerden und Nachweis der weiteren Umstände kam.

Der Indizienbeweis kann bei maschinellen Entscheidungen jedoch wesentlich erschwert sein, da es häufig an Kontextinformationen fehlt, aus denen sich solche Indizien ergeben können. Wenn Informationen über den Betroffenen lediglich anhand eines Eingabeformulars einer Website erhoben werden und diesem sodann lediglich ein Ergebnis mitgeteilt wird, fehlt es jedoch am weiteren Kontakt, aus dem sich Indizien für eine Diskriminierung ergeben können.

7.2.3.2 Beweislastumkehr

Ein wesentliches Mittel des Rechtsschutzes gegen Diskriminierung sind Beweisregeln, die insbesondere für zivilrechtliche Ansprüche und im Zivilprozess von Bedeutung sind. Das stärkste Mittel ist insoweit die Beweislastumkehr. Ein Beispiel für eine solche Regelung ist etwa § 22 AGG, der bei Vorliegen von Indizien für eine Benachteiligung eine Beweislastumkehr anordnet.

7.2.3.2.1 Vorschläge zur Einführung einer Beweislastumkehr

Das Mittel der Beweislastumkehr erscheint attraktiv, um die Schwierigkeit der Feststellung einer Diskriminierung zu umgehen.

Entsprechend schlägt etwa Martini für Haftungsprozesse eine Beweislastumkehr vor, die bei persönlichkeitssensiblen Softwareanwendungen greifen soll, wenn der Nutzer Tatsachen vorträgt, die „mit überwiegender Wahrscheinlichkeit darauf schließen lassen, dass unzulässige Parameter Eingang in die Entscheidung gefunden haben“.[292]

Ob Martini eine Beweislastumkehr im engeren Sinne oder eine Beweiserleichterung nach Art des Anscheinsbeweises meint, bleibt letztlich offen, da im Weiteren davon gesprochen wird, der Anbieter der Software müsse sich durch „Erschütterung einer Kausalitätsvermutung“ freizeichnen.[293]

7.2.3.2.2 Herausforderungen einer gesetzlichen Beweislastumkehr

Eine solche Beweislastregel, die *de lege lata* wohl nicht begründet werden kann, müsste durch den Gesetzgeber eingeführt und präzisiert werden. Dabei sind folgende Aspekte von Bedeutung:

7.2.3.2.2.1 Anwendungsbereich der Beweislastregel

Eine wesentliche Frage für eine solche Beweislastumkehr ist deren Anwendungsbereich, konkret die Frage, ob die Regel nach dem Vorbild des AGG auf bestimmte, enumerativ zu bezeichnende Bereiche beschränkt sein soll oder, wie es offenbar Martini vorschwebt, einen allgemeinen Anwendungsbereich haben soll. Dabei erscheint der von Martini

[291] Vgl. OLG Köln, Urteil v. 19.01.2009 – 24 U 51/09 –, NJW 2010, 1676, 1677.
[292] Martini, JZ 2017, 1017 ff.
[293] Martini, JZ 2017, 1017, 1024.

vorgeschlagene Bereich der „persönlichkeitssensiblen Softwareanwendungen“ zu unpräzise, um die erforderliche Rechtsklarheit zu gewährleisten.

Die Frage nach dem angemessenen Anwendungsbereich einer solchen Regel kann im Rahmen dieser Studie nicht geklärt werden. Es erscheint angesichts der Bedeutung einer Beweislastumkehr jedoch geboten, insoweit eine klare Regelung zu finden.

7.2.3.2.2.2 Voraussetzungen der Beweislastumkehr

Als Voraussetzung der Beweislastumkehr wird im AGG auf die Vermutung einer Benachteiligung aufgrund von Indizien abgestellt. Martini verweist, in der Sache wohl nicht anders, auf die „überwiegende Wahrscheinlichkeit“ der Verwendung eines unzulässigen Parameters. Eine konkretere Beschreibung der Voraussetzungen wird angesichts der Vielzahl möglicher Sachverhaltskonstellationen wohl auch nicht möglich sein.

7.2.3.2.2.3 Möglichkeit des Gegenbeweises

Rechtsfolge der Beweislastumkehr ist, dass der Gegner, in der Regel die Partei, die eine maschinelle Beurteilung verwendet, den Gegenbeweis für das Nichtvorliegen einer Diskriminierung erbringen muss. Gegenbeweis meint insoweit den vollen Beweis, also den Ausschluss der Diskriminierung zur vollen Überzeugung (Fehlen vernünftiger Zweifel) des Gerichts.

Die Bedeutung der Beweislastumkehr hängt daher nicht zuletzt von der Möglichkeit eines solchen Gegenbeweises ab.

Martini verweist insoweit auf die Vorlage protokollierter Programmabläufe.[294] Damit bestehen dieselben oder ähnliche Anforderungen, wie sie im Rahmen der Code-Analyse zur Feststellung einer Diskriminierung existieren, lediglich mit dem umgekehrten Vorzeichen, dem Nichtbestehen einer Diskriminierung. Daher gibt es letztlich dieselben Schwierigkeiten, d.h. Aufwand, Undurchführbarkeit und Offenlegung. Dem Gesichtspunkt der Offenlegung tritt Martini mit der Möglichkeit einer Offenlegung in camera, also ausschließlich gegenüber dem Gericht, gegenüber.

7.2.3.2.3 Rechtsfolgen bei Diskriminierungsverdacht

Eine wohl eher theoretische Möglichkeit besteht darin, materielle Rechtsfolgen, etwa Schadensersatzansprüche, bei einem bloßen Diskriminierungsverdacht greifen zu lassen, so dass der schwierige Nachweis einer tatsächlichen Diskriminierung nicht erforderlich wäre.

Derartige Ansätze sind offensichtlich nicht überzeugend. Sie entsprechen der Sache nach den Rechtsfolgen einer unwiderleglichen Vermutung für Diskriminierung und schneiden sogar die Möglichkeit des Gegenbeweises ab, ohne dass hierfür eine Legitimation besteht.

7.2.3.3 Ermöglichung der Feststellung von Diskriminierung in maschinellen Entscheidungen

Ein konzeptionell völlig anderer Weg lässt sich durch Maßnahmen beschreiten, die die Feststellung von Diskriminierung durch maschinelle Entscheidungen ermöglichen und gegebenenfalls gegen den Willen des Betreibers eines ADM-Verfahrens durchsetzen. Diese Konzeption ist mit der oben (3.4.5) genannten Frage nach dem „Recht auf Analyse“

[294] Martini, JZ 2017, 1017, 1024.

angesprochen. Entsprechend soll nachfolgend am Beispiel der Code-Analyse sowie des Testens von ADM-Systemen untersucht werden, wie die Durchführung dieser Maßnahmen rechtlich gesichert werden kann.

7.2.3.3.1 Durchsetzung einer Code-Analyse

Um die Code-Analyse zu ermöglichen, sind sehr unterschiedliche legislative Maßnahmen denkbar, je nach dem Kontext, in dem sie relevant werden kann.

Soweit eine behördliche Aufsicht über maschinelle Entscheidungen besteht, kann der Gesetzgeber der zuständigen Behörde Befugnisse zur Einsicht und Untersuchung der Software gewähren. Da insoweit seitens der Behörde Verschwiegenheitspflichten bestehen, kann der Einwand der Verletzung von Geschäftsgeheimnissen entkräftet werden. Es bleibt freilich das Problem des Aufwandes, der mit der Code-Analyse verbunden ist.

Im Verhältnis zwischen Privaten ist die Einräumung von Einsichts- und Untersuchungsrechten oder entsprechenden Pflichten oder Lasten auf Seiten des Betreibers des Systems problematisch, da Geschäftsgeheimnisse offenbart werden müssen.

7.2.3.3.2 Durchsetzung von Tests maschineller Entscheidungsverfahren

Die Ermöglichung von Tests oder der Mitwirkung von Tests maschineller Entscheidungen kann wiederum durch unterschiedliche Maßnahmen erfolgen. Diese hängen entscheidend von der rechtlichen Einbettung des Tests ab: So sind etwa im Bereich einer behördlichen Aufsicht Untersuchungsbefugnisse und entsprechende Mitwirkungspflichten von Bedeutung. Dagegen stellen sich im Verhältnis zwischen Privaten Fragen nach einem Anspruch auf Testdurchführung oder Mitwirkungspflichten.

Die rechtliche Sicherung der tatsächlichen Durchführung von Tests ist für den Rechtsrahmen von Tests von wesentlicher Bedeutung und soll daher im Einzelnen (siehe unten Kapitel 7.3.3.) erörtert werden.

7.3 Rechtlicher Rahmen von Testverfahren für ADM-Systeme

7.3.1 Die Bedeutung von Tests für die Kontrolle von ADM-Systemen und algorithmischen Entscheidungen

Als ein wesentliches Ergebnis der Studie hat sich ergeben, dass Tests ein zentrales Mittel für die Kontrolle algorithmischer Entscheidungen sind. Entsprechend sollte dieses Instrument für die rechtliche Regulierung algorithmischer Entscheidungen fruchtbar gemacht werden.

Aus rechtlicher Sicht stellen sich damit zahlreiche Fragen; zentral sind insbesondere folgende Aspekte:

- **Rechtliche Bedeutung (Rechtsfolgen) von Tests:** Wichtig ist, welche rechtliche Bedeutung Tests und Testergebnisse für ADM-Verwender haben, konkret welche Rechtsfolgen sich aus der Durchführung oder auch der Nichtdurchführung von Tests und aus Testergebnissen ableiten lassen.

- **Durchsetzung der Durchführung von Tests:** Ein wesentlicher Aspekt betrifft die Durchführung von Tests im weiteren Sinne. Neben den Anforderungen an das Testverfahren etc. (siehe unten) ist es von besonderem Interesse, ob Ansprüche Privater oder Ermächtigungsgrundlagen von Behörden zur Durchführung von Tests gegen den Willen des Verwenders des ADM-Systems geschaffen werden sollen, sowie flankierende Maßnahmen und Regelungen wie Auskunftsansprüche, Kostentragungsregeln etc.
- **Anforderungen an Testverfahren:** Soweit Tests Rechtsfolgen beigemessen werden sollen, ist es von Bedeutung, welche Anforderungen an Testverfahren zu stellen sind. Diese umfassen sowohl materielle Anforderungen an den im Test verwendeten Prüfstandard als auch verfahrensmäßige Anforderungen an die Durchführung von Tests.

Da das Potenzial von Tests für die Regulierung von ADM-Systemen sowie von algorithmischen Entscheidungen entscheidend von diesen Aspekten abhängt, werden diese nachfolgend erörtert.

7.3.2 Rechtliche Bedeutung von Tests

Die rechtliche Relevanz von Testverfahren für ADM-Systeme und Tests algorithmischer Entscheidungen kann sich auf sehr unterschiedliche Weise ergeben. Entsprechend sind ganz unterschiedliche Systematisierungen der rechtlichen Bedeutung möglich. Für die Zwecke dieser Studie sollen folgende Fallgruppen unterschieden werden:

- Rechtsfolgen können bereits an die Durchführung oder Nichtdurchführung eines Tests, meist im Zusammenhang mit einem bestimmten Testergebnis, geknüpft werden, etwa wenn die Durchführung eines Tests (einschließlich eines positiven Ergebnisses) eine rechtliche Voraussetzung für den Einsatz eines ADM-Systems ist.
- Rechtsfolgen können sich aus einer Tatsache ergeben, die mit einem bestimmten Testergebnis festgestellt wird, etwa wenn sich durch einen Test feststellen lässt, dass bei einer algorithmischen Entscheidung keine Diskriminierung vorlag.

7.3.2.1 Durchführung von Tests und Existenz von Testergebnissen

7.3.2.1.1 Durchführung von Tests im Recht

Die Durchführung oder Nichtdurchführung eines Tests, meist im Zusammenhang mit einem bestimmten Testergebnis, kann in ganz unterschiedlicher Weise Rechtsfolgen auslösen.

Eine wesentliche Fallgruppe ist die Durchführung eines Tests zur Erfüllung eines rechtlichen Prüfungserfordernisses. Diese Anforderungen können in unterschiedlicher Weise bestehen, etwa als Zulassungsvoraussetzung für den Einsatz eines ADM-Systems, als von einer Aufsichtsbehörde angeordnet oder im Rahmen einer vertraglichen Beziehung vereinbarte Maßnahme. Ähnliches gilt, wenn durch Gesetz eine Prüfung oder Zertifizierung angeordnet wird.

In diesen Fällen wird durch die Durchführung des Texts als solche, meist in Verbindung mit einem positiven Testergebnis, eine rechtliche Verpflichtung erfüllt, die diese Prüfung konkret verlangt.

Die Durchführung eines Tests kann darüber hinaus Bedeutung haben im Rahmen von Sorgfaltspflichten, etwa im Zusammenhang mit zivilrechtlichen Haftungsansprüchen oder bei der Verhängung von Sanktionen. Es ist beispielsweise offensichtlich, dass die Durchführung oder Nichtdurchführung eines Tests von ADM-Systemen entscheidend für ein etwaiges Verschulden in Bezug auf Diskriminierung oder andere Fehler algorithmischer Entscheidungen oder den Einsatz eines fehlerhaften ADM-Systems sein kann.

Die Vorlage oder Existenz eines bestimmten Testergebnisses kann weiterhin rechtliche Pflichten auslösen. Dies gilt insbesondere bei negativen Testergebnissen. Wenn etwa dem Betreiber eines ADM-Systems ein Testergebnis vorgelegt oder bekannt wird, das die Existenz eines Fehlers (z.B. Diskriminierung) nachweist oder den Verdacht auf einen Fehler begründet, können sich aus diesem Umstand Handlungspflichten ergeben, konkret die Pflicht zu eigenen Nachforschungen, oder etwa Meldepflichten gegenüber einer Behörde etc.

Die Durchführung von Tests ist schließlich eng verknüpft mit der Selbstregulierung insbesondere durch Kodizes und Zertifizierung. In beiden Fällen werden vertrauenswürdige Verfahren der Selbstregulierung die Durchführung von Tests der ADM-Systeme vorsehen. Das Vertrauen in die Befolgung eines Kodex oder die Vergabe eines Zertifikats beruht typischerweise entscheidend auf der Durchführung eines geeigneten Tests durch eine vertrauenswürdige Institution. Daher kann es sinnvoll oder gar geboten sein, das Testverfahren rechtlich zu regeln, um dieses Vertrauen zu schützen.

Damit ergibt sich Folgendes: Die Durchführung eines Tests und das Bestehen eines Testergebnisses können, wie die Darstellung gezeigt hat, rechtssystematisch ganz unterschiedlich eingebettet sein: Insbesondere können sie einerseits zur Erfüllung einer rechtlichen Pflicht dienen, andererseits aber auch rechtliche Pflichten auslösen. Darüber hinaus können sie tatsächliche Folgen (wie Vertrauen) auslösen, die als solche gegebenenfalls nicht Gegenstand rechtlicher Regelung sein müssen, aber gegebenenfalls in ihrem Tatbestand, konkret dem Testverfahren, Gegenstand rechtlicher Regelung sein können.

Diese unterschiedliche Einbettung in die Rechtsordnung ist wesentlich für die Ausgestaltung der rechtlichen Rahmenbedingungen von Testverfahren für ADM-Systeme und algorithmische Entscheidungen.

Damit ergeben sich Aufgaben und Handlungsoptionen für den Gesetzgeber auf mehreren Ebenen.

7.3.2.1.2 Klärung der Rechtsfolgen von Testdurchführung und Testergebnissen

Die Durchführung, meist im Zusammenhang mit einem positiven Testergebnis, oder das Unterlassen eines Tests kann wie dargestellt, in unterschiedlicher Weise Rechtsfolgen auslösen. Dabei sind insbesondere die Durchführung von Tests zur Erfüllung einer rechtlichen Verpflichtung, die Bedeutung im Zusammenhang mit einem Verschulden oder die Beweisführung sowie Selbstregulierung von Interesse.

Soweit die Durchführung oder Nichtdurchführung eines Tests für das Verschulden von Relevanz ist, hängt dies eng mit dem Bestehen einer rechtlichen Pflicht zusammen, da zumindest bei Missachtung einer rechtlichen Pflicht zur Durchführung regelmäßig auch ein Verschulden in Form der Verletzung einer Sorgfaltspflicht vorliegt. Daher konzentrieren sich die nachfolgenden Überlegungen auf den Aspekt der Erfüllung einer rechtlichen Pflicht.

7.3.2.1.3 Fehlen klarer gesetzlicher Pflichten zum Testen von ADM-Systemen

Rechtliche Pflichten zur Durchführung eines Tests von ADM-Systemen können sich in allen Rechtsgebieten, insbesondere aus zivilrechtlichen Sorgfaltsanforderungen sowie aus öffentlich-rechtlichen Pflichten, ergeben. Sie können auf expliziter gesetzlicher Anordnung beruhen oder sich aus allgemeinen Grundsätzen, vor allem aus Generalklauseln zur Sorgfalt, ableiten lassen.

Nach derzeitiger Rechtslage bestehen soweit erkennbar keine bzw. allenfalls punktuell ausdrückliche Regeln zu Tests. Auch § 28b BDSG a.F. und § 31 BDSG n.F. enthalten eine solche Pflicht nicht ausdrücklich.

Rechtliche Pflichten zum Testen von ADM-Systemen bestehen daher nach derzeitiger Rechtslage nur, soweit sie sich aus allgemeinen Normen ableiten lassen. Dies dürfte in vielen Fällen durchaus möglich sein. Ob sich aus § 31 BDSG n.F. eine solche Pflicht ableiten lässt, ist fraglich.

Es dürfte sich aber aus § 823 BGB eine Pflicht zur Durchführung von Tests eines ADM-Systems vor Inbetriebnahme des Systems ableiten lassen, soweit etwa eine Persönlichkeitsrechtsverletzung durch ein fehlerhaftes ADM-System droht.

Der Anwendungsbereich und der konkrete Inhalt einer solchen aus der Generalklausel des Deliktsrechts ableitbaren Pflicht sind jedoch derzeit nicht absehbar. So ist schon unklar, ob sie sich auf den Hersteller oder den Nutzer eines solchen Systems (oder beide) bezieht. Vor allem ist nicht absehbar, in welchen Fällen ein solches ADM-System, für das eine Testpflicht gilt, vorliegt. Im Fall eines ADM-Systems zur Unterstützung der Personalauswahl ist schon unklar, in welchen Fällen ein Fehler zu einer Persönlichkeitsrechtsverletzung führt. Insbesondere sind die Anforderungen an den Inhalt des Testverfahrens derzeit nicht geklärt. Diese Unklarheit ist für eine breite Durchsetzung der Tests von ADM-Systemen hinderlich.

Es ist davon auszugehen, dass sich aus der Praxis, speziell der Rechtsprechung, in Zukunft eine Klärung in Bezug auf das Bestehen der Testpflicht ergibt.

Jedoch sollte der Gesetzgeber insoweit nicht abwarten, da der Klärungsprozess durch die Rechtsprechung etliche Jahre in Anspruch nehmen kann. Vielmehr sollte der Gesetzgeber aktiv werden.

7.3.2.1.4 Forschungsbedarf

Zu klären ist im Besonderen das Gefahrenpotenzial von ADM-Systemen, etwa für Persönlichkeitsrechte der betroffenen Personen, das in dieser Studie nur angerissen werden konnte. Dabei wird sich voraussichtlich zeigen, dass jedenfalls in einigen Bereichen erhebliche Nachteile drohen.

Zu klären ist auch die Verfügbarkeit und Leistungsfähigkeit von Testverfahren zu ADM-Systemen und algorithmischen Entscheidungen.

7.3.2.1.5 Einführung gesetzlicher Prüfpflichten

Auf der Grundlage der Forschung ist zu entscheiden, ob und in welchen Bereichen gesetzliche Prüfpflichten für ADM-Systeme und algorithmische Entscheidungen gesetzlich geregelt werden sollten.

Anhand der Ergebnisse der Studie lassen sich folgende Prognosen wagen:

1. Eine Pflicht zur manuellen Überprüfung einer einzelnen algorithmischen Entscheidung ist nur in engen Teilbereichen sinnvoll. Ob über Art. 22 DSGVO hinaus eine Regelung geboten oder möglich ist, ist derzeit nicht abzusehen.

2. Eine Pflicht zur Durchführung von Tests von ADM-Systemen vor deren Einsatz sowie begleitend zu deren Einsatz als Teil der laufenden Überwachung wird in vielen Bereichen geboten sein. Dies wird stets der Fall sein, soweit ADM-Systeme in für den Betroffenen wesentlichen Bereichen eingesetzt werden. Die Personalauswahl gehört hierzu ebenso wie Versicherungstarife, namentlich im Bereich der Krankenversicherung.

Soweit geeignete Testverfahren allgemein verfügbar sind, wird die Durchführung eines Tests zumutbar sein. Damit die Voraussetzungen einer Pflicht, die sich aus Generalklauseln ableiten ließe, vorliegen, sollte die Pflicht zur Vermeidung von Rechtsunklarheit und Rechtsunsicherheit ausdrücklich geregelt werden (dazu unten Kapitel 8.3.3.2).

7.3.2.2 Testergebnis und Schluss auf rechtserhebliche Tatsachen

7.3.2.2.1 Fallgruppen

Eine weitere wesentliche Fallgruppe der rechtlichen Bedeutung von Tests ergibt sich durch die Möglichkeit, von einem Testergebnis auf das Vorliegen einer rechtlich erheblichen Tatsache zu schließen.

Wenn sich beispielsweise aufgrund eines Tests einer konkreten algorithmischen Entscheidung ergibt, dass bei der konkreten Entscheidung eine Diskriminierung vorlag, kann unter Umständen das Testergebnis dazu führen, dass rechtlich von der Existenz einer Diskriminierung auszugehen ist.

Der damit angesprochene Schluss von der Existenz eines Testergebnisses auf die Existenz einer rechtlich erheblichen Tatsache ist in sehr unterschiedlichen rechtlichen Konstellationen relevant und wirft zahlreiche Fragen in verschiedenen Rechtsgebieten auf. So hat dieser Schluss etwa für den Betreiber eines ADM-Systems, den Betroffenen, eine Aufsichtsbehörde oder ein Gericht Bedeutung. Insbesondere stellt sich die Frage, unter welchen Voraussetzungen der Schluss gezogen werden kann oder gezogen werden muss, wann also der Test zweifelsfrei auf diese Tatsache weist. Gleichzeitig ist relevant, welche Bedeutung ein Verdacht oder das Naheliegen der Tatsache unterhalb der Schwelle voller Überzeugung für die Existenz der jeweiligen Tatsache hat.

Diese überaus komplexe Fragestellung kann im Rahmen dieser Studie nicht untersucht werden. Da dieser Schluss aber für weitere Fragestellungen der Untersuchung, etwa legislative Maßnahmen, von entscheidender Bedeutung ist, sollen zur Vereinfachung zwei Annahmen getroffen werden:

1. Annahme: Es wird zahlreiche Sachverhalte geben, in denen aufgrund eines Testergebnisses über eine algorithmische Entscheidung ein Entscheidungsträger, etwa ein Gericht, zu der vollen Überzeugung gelangt, dass die Entscheidung an einem rechtlich relevanten Fehler leidet, etwa eine (rechtlich relevante) Diskriminierung enthält.

2. Annahme: Es wird zahlreiche Sachverhalte geben, in denen es aufgrund eines Testergebnisses über eine algorithmische Entscheidung aus Sicht des Entscheidungsträgers naheliegt, dass die Entscheidung an einem rechtlich relevanten

Fehler leidet, etwa eine (rechtlich relevante) Diskriminierung enthält, dies aber nicht ohne weiteres zweifelsfrei auf die Existenz eines solchen Fehlers schließen lässt.

Mit diesen Annahmen ist insbesondere die beweisrechtliche Relevanz von Testverfahren angesprochen. In der erstgenannten Konstellation wird der Entscheidungsträger aufgrund des Testergebnisses auf die jeweilige Tatsache, hier auf die Existenz eines Fehlers der algorithmischen Entscheidung, schließen.

Die Annahmen beziehen sich zur Vereinfachung auf den Test einer konkreten algorithmischen Entscheidung. In aller Regel hat das Testergebnis rechtliche Auswirkungen auf das betroffene ADM-Verfahren insgesamt. So wird, wenn das ADM-System unverändert eingesetzt wird, regelmäßig davon auszugehen sein, dass das System insgesamt fehlerhaft ist. Damit werden Pflichten beim Einsatz des Verfahrens ausgelöst.

In Bezug auf die rechtliche Bedeutung von Testergebnissen sind zwei Fallgruppen nach dem Gegenstand des Testergebnisses zu unterscheiden: die Überprüfung einer konkreten algorithmischen Entscheidung zum einen und die Überprüfung eines ADM-Systems zum anderen.

7.3.2.3.1 Überprüfung einer algorithmischen Entscheidung

Tests an ADM-Systemen können von Bedeutung sein, wenn eine konkrete algorithmische Entscheidung überprüft wird, etwa im Rahmen eines gerichtlichen Verfahrens wegen einer Diskriminierung oder einer fehlerhaften Auswahlentscheidung.

Die Bedeutung der durch Tests an ADM-Systemen erzielten Ergebnisse im Rahmen der Überprüfung konkreter algorithmischer Entscheidungen folgt nach geltendem Recht den allgemeinen Regeln. Wenn etwa aufgrund eines Testergebnisses festgestellt wird, dass bei einer algorithmischen Entscheidung eine Diskriminierung vorlag, treten die allgemeinen Rechtsfolgen einer Diskriminierung ein. Die Überprüfung algorithmischer Entscheidungen auf Fehler, etwa Diskriminierung, fügt sich damit nahtlos in das geltende Recht ein.

Ob sich daraus ableiten lässt, dass insoweit kein Gesetzgebungsbedarf besteht, ist jedoch zweifelhaft. Zum einen dürften sich im Zusammenhang mit ADM-Systemen und algorithmischen Entscheidungen durchaus rechtlich relevante Probleme ergeben, die bei menschlichen Beurteilungen hingenommen werden, im Zusammenhang mit manueller Erledigung jedoch rechtlichen Regelungen zugänglich werden. Zum anderen ist damit zu rechnen, dass im geltenden Recht Lücken oder Unklarheiten bestehen.

Darauf folgt zunächst erheblicher Forschungsbedarf in Bezug auf Fehler durch algorithmische Entscheidungen. Gegenstand der Forschung sollten unter anderem folgende Aspekte sein:

7.3.2.3.2 Vorliegen und Rechtsfolgen des Fehlers einer algorithmischen Entscheidung

Zu klären ist, unter welchen Voraussetzungen im Fall einer algorithmischen Entscheidung ein rechtlich relevanter Fehler vorliegt, der Rechtsfolgen auslösen soll. Wenn beispielsweise eine Überprüfung einer algorithmischen Entscheidung ergibt, dass mit hoher Wahrscheinlichkeit der ausländisch klingende Name eines Verbrauchers im Zusammenhang mit dessen Wohnanschrift zu einer negativen Kreditentscheidung geführt hat, so ist nach geltendem Recht unklar, ob und welche Rechtsfolgen einer Diskriminierung durch diesen Befund ausgelöst werden.

Ein Aspekt ist etwa, unter welchen Voraussetzungen eine Diskriminierung, die sich aufgrund eines Tests ergibt, erheblich ist. Dabei wird zu beachten sein, dass Tests maschineller Entscheidungen deutlich eher zur Feststellung oder zum Verdacht einer Diskriminierung führen, als es angesichts der genannten Feststellungschwierigkeiten bei menschlichen Entscheidungen der Fall ist. Daher sind auch die Rechtsfolgen eines Fehlers gegebenenfalls neu zu bestimmen, etwa durch ein abgestuftes System, das nicht bei jeglichem Fehler zur Unwirksamkeit oder Anfechtbarkeit der Entscheidung führt oder Schadensersatzansprüche auslöst.

7.3.2.3.3 Gap-Analyse des Rechtsrahmens hinsichtlich algorithmischer Entscheidungen

Grundlegender Forschungsbedarf ergibt sich hinsichtlich potentieller Lücken im Rechtsrahmen bezüglich Fehlern algorithmischer Entscheidungen. Wie dargestellt, haben zentrale Instrumente des Schutzes gegen Diskriminierungen oder andere Fehler von Entscheidungen, insbesondere das AGG hinsichtlich Diskriminierung, sowie dem Datenschutzrecht (Art. 22 DSGVO, § 31 BDSG), zwar durchaus einen breiten Anwendungsbereich, decken aber nur Teilaspekte ab. So sichert Art. 22 DSGVO zwar eine menschliche Mitwirkung bei der Entscheidung, enthält aber keine spezifischen Regeln zu Fehlern algorithmischer Entscheidungen. Daneben wird Rechtsschutz auch durch Generalklauseln des Zivilrechts gewährleistet. Jedoch ist derzeit nicht klar festzustellen, ob die Schutzinstrumente einen hinreichenden Schutz gegenüber ADM-Systemen erbringen und inwieweit Lücken bestehen. Die Analyse etwaige Lücken wird derzeit erschwert durch die Existenz zahlreicher ungeklärter Grundsatzfragen, die im Rahmen dieser Studie angesprochen wurden.

Es erscheint daher sinnvoll, vor Veranlassung legislativer Maßnahmen im Rahmen der erforderlichen Forschung eine Gap-Analyse des geltenden Rechts zu zentralen Fehlern algorithmischer Entscheidungen, insbesondere Diskriminierung, durchzuführen.

7.3.2.3.4 Überprüfung von ADM-Systemen

Tests an ADM-Systemen können rechtlich ebenso von Relevanz sein, wenn es nicht um die Überprüfung einer algorithmischen Entscheidung geht, sondern um die Qualität der Entscheidungen und damit die Funktionsfähigkeit des Systems insgesamt. In diesem Fall hat der Test oft über den Einzelfall hinaus Bedeutung.

Derartige Fragen können sich in zivilrechtlichen Kontexten stellen. Vor allem können Testergebnisse im Rahmen kollektivrechtlicher oder aufsichtsrechtlicher Verfahren von Bedeutung sein. Für die nachfolgenden Überlegungen wird die Situation zugrunde gelegt, dass ein Test an einem ADM-System durchgeführt wurde und dieser Test auf einen Fehler des Systems, etwa diskriminierende Entscheidungen, hinweist.

7.3.2.3.5 Bewertung von Testergebnissen im Zivilrecht

Die Frage nach der Bedeutung eines Testergebnisses für die Bewertung eines ADM-Systems kann in den unterschiedlichsten zivilrechtlichen Kontexten von Bedeutung sein, etwa im Zusammenhang mit der Mangelhaftigkeit des Systems, die im Verhältnis zwischen Käufer und Verkäufer des Systems relevant ist. Sie kann auch Bedeutung haben zur Auslösung von Verhaltenspflichten, insbesondere für die Pflicht, die Verwendung eines ADM-Systems zu unterlassen. Eine solche Unterlassungspflicht kann sich etwa aus den allgemeinen Regeln des Vertrags- und Deliktsrechts ergeben. So wird sich aus § 823 BGB, der Generalklausel des Deliktsrechts, sowie aus § 1004 BGB (analog) wohl ableiten lassen,

dass der Verwender eines ADM-Systems, dem aufgrund eines Testergebnisses bekannt ist, dass das von ihm eingesetzte System fehlerhafte Entscheidungen trifft, dieses nicht weiterhin nutzen darf, soweit beim Einsatz des Systems Rechte Dritter, etwa Persönlichkeitsrechte, verletzt werden. Insoweit bestehen ein Unterlassungsanspruch des Betroffenen (§ 1004 analog) sowie, wenn ein Schaden eingetreten ist, ein Schadensersatzanspruch aus § 823 BGB. In Schuldverhältnissen ergeben sich entsprechende Pflichten als vertragliche Nebenpflicht nach § 241 BGB, gegebenenfalls in Verbindung mit § 311 BGB, bei deren Verletzung ein Schadensersatzanspruch nach § 280 BGB entsteht. Es ist jedoch offensichtlich, dass der Schutz vor fehlerhaften algorithmischen Entscheidungen auf dieser gesetzlichen Grundlage jedenfalls derzeit recht schwach ist. Dies beruht schon darauf, dass ein zivilrechtlicher Unterlassungs- oder Schadensersatzanspruch eines Betroffenen nur im Einzelfall wirkt, nicht aber die Verwendung gegenüber Dritten betrifft. Es liegt nahe, dass die Ergänzung durch Aufsicht oder Klagen von Verbänden notwendig sein wird.

Ähnliche Fragen ergeben sich für den Hersteller von ADM-Systemen. Dieser unterliegt jedenfalls den allgemeinen Regeln des Delikts- und des Produkthaftungsrechts. Hier besteht derzeit das Problem, dass die Anwendbarkeit des ProdHaftG auf reine Software durchaus noch umstritten ist. Insoweit ist der europäische Gesetzgeber bei der anstehenden Überarbeitung der Produkthaftungsrichtlinie gefordert, für eine Klarstellung/Klärung zu sorgen.

Es dürfte nicht zweifelhaft sein, dass sich aus den Regeln der deliktischen Produzentenhaftung nach § 823 BGB auch Pflichten für den Hersteller von ADM-Systemen ergeben, soweit diese geeignet sind, Persönlichkeitsrechte zu verletzen. Es dürfte auch nicht fraglich sein, dass ein Hersteller insoweit verpflichtet ist, das System angemessen zu testen, bevor es in Verkehr gebracht wird. Auch gilt die Produktbeobachtungspflicht, wonach der Hersteller verpflichtet ist, das ADM-System auch nach Inverkehrbringen in Bezug auf die Gefahr der Verletzung von Persönlichkeitsrechten zu beobachten und zu reagieren. Ein negatives Testergebnis löst daher zumindest die Pflicht aus, diese zu bewerten und gegebenenfalls Folgen für den Einsatz des Systems abzuleiten. Daraus könnten Informationspflichten gegenüber den Verwendern des Systems entstehen und gegebenenfalls sogar die Pflicht, das System zurückzurufen (Produktrückruf).

Allerdings bestehen erhebliche Unklarheiten über die Konkretisierung dieser Pflichten auf ADM-Systeme. Bei Systemen, die *Machine Learning* einsetzen, kommen weitere grundlegende Fragen hinzu. So ist derzeit jedenfalls noch völlig ungeklärt, wie weit die Pflicht der Hersteller reicht, insbesondere wenn das Trainieren des Systems oder die Datengrundlage beim Einsatz des Systems im Verantwortungsbereich des Verwenders des Systems liegt.

7.3.2.3.6 Testergebnisse im Aufsichtsrecht

Der Rechtsrahmen der Überprüfung von ADM-Systemen im Rahmen der Aufsicht ist äußert unklar, da es an einem spezifischen aufsichtsrechtlichen Rahmen fehlt. So sind Aufsichtsmaßnahmen im Rahmen der Finanzaufsicht denkbar, wenn ein ADM-System im Finanzbereich eingesetzt wird, ebenso nach dem Datenschutzrecht. In beiden Beispielsfällen gibt es keine Regelung zu Tests oder zur Bedeutung von Testergebnissen von ADM-Systemen.

Entsprechend ist auch die Bedeutung eines konkreten Testergebnisses unklar. Wenn etwa einer Datenschutzaufsichtsbehörde vorgetragen wird, ein Test habe ergeben, dass ein

ADM-System diskriminierende Entscheidungen vorschlage, so ist in vielerlei Hinsicht unklar, welche Befugnisse die Behörde insoweit hat oder ob sie gar eingreifen muss. In welchem Maße die inhaltlich richtige Verwendung von Daten – darum geht es in diesem Fall – Schutzgegenstand des Datenschutzrechts ist, ist durchaus noch nicht geklärt. Vor allem wird sich für die Aufsicht die Frage stellen, unter welchen Voraussetzungen sie von einem Testergebnis auf einen grundsätzlichen Mangel des Systems schließen kann, da wohl nur dann eine Untersagungsverfügung gegen den Betreiber des Systems in Betracht kommt. Ähnliche Fragen stellen sich bei der Finanzaufsicht.

Aufsichtsrechtliche Maßnahmen gegen Hersteller von ADM-Systemen kommen nach geltendem Recht in den hier betrachteten Fallgruppen wohl nicht in Betracht, so dass sich die Frage nach der Bedeutung von Testergebnissen nicht stellen dürfte.

7.3.2.3.7 Klärungsbedarf zur Bedeutung von Testergebnissen für die Herstellung und Verwendung von ADM-Systemen

Die Überlegungen zeigten, dass sich nach geltendem Recht schwierige Fragen auch über die Bewertung eines Testergebnisses und seiner Relevanz für die Herstellung und Verwendung von ADM-Systemen ergeben.

Wie dargestellt, ist derzeit durchaus nicht klar, ob und unter welchen Voraussetzungen ein ADM-System im Fall eines negativen Testergebnisses noch eingesetzt werden darf. Dies hängt jedenfalls teilweise von der Unklarheit über die Qualität von Testverfahren und darauf ableitbaren Testergebnissen ab.

Soweit Privatpersonen Rechte geltend machen wollten, ist unklar, inwieweit überhaupt Zugang zu den Informationen besteht, die eine Bewertung des Testergebnisses und seiner Folgen ermöglichen.

Ähnliches gilt im Fall der Aufsicht, die zwar regelmäßig eigene Untersuchungsbefugnisse hat, also selbst Tests durchführen könnte, aber regelmäßig keine spezifischen Kompetenzen zur Bewertung von Tests hat. Dies führt zu erheblichen praktischen Schwierigkeiten: Wenn etwa einer Datenschutzaufsichtsbehörde oder der Finanzaufsicht vorgetragen wird, ein ADM-System entscheide diskriminierend, und dabei auf ein Testergebnis verwiesen wird, so wird es für die Behörde aus eigener Kompetenz kaum möglich sein, das Testergebnis und den entsprechenden Vortrag zu bewerten. Es müsste also auswärtige Kompetenz eingeholt werden, wofür aber typischerweise keine entsprechenden Ressourcen bereitstehen. In jedem Fall entsteht für die Behörde ein sehr erheblicher Aufwand für die Einarbeitung in die spezifischen Aspekte der ADM-Systeme, die nicht zum „Kerngeschäft“ der Behörde gehören. Da in den meisten Aufsichtsbereichen, jedenfalls im Bereich der Datenschutz- und Finanzaufsicht, eine Vielzahl anderer Aufgaben vorliegt, ist damit zu rechnen, dass ein weiteres Tätigwerden unterbleibt, soweit nicht von dritter Seite klare Aussagen zur Qualität des Testergebnisses vorgelegt werden.

Auch bei der Bewertung von Testergebnissen bestehen erhebliche Schwierigkeiten, da die Aussagekraft eines Tests entscheidend von der Qualität des Teststandards und des Testverfahrens abhängt. Insoweit fehlt es jedoch an allgemein anerkannten Maßstäben. Dies ist sowohl für zivilrechtliche Mechanismen, etwa Produkthaftung, als auch für die Aufsicht von Bedeutung, da in beiden Fällen Rechtsklarheit über die Bewertung eines Testergebnisses entscheidend für die Frage ist, ob zivilrechtliche Rechte geltend gemacht werden oder Maßnahmen der Aufsicht ergriffen werden. So wird ein Betroffener keine Rechte wegen Verletzung einer Produktbeobachtungspflicht geltend machen, wenn nicht

klar ist, welche Bedeutung ein etwaiges negatives Testergebnis hat. Ähnliches gilt im Fall der Aufsicht.

Für jeglichen Rechtsschutz in Bezug auf ADM-Systeme wird es daher wichtig sein, dass Aufsichtsbehörden, gegebenenfalls auch Betroffene, Zugang zu Testergebnissen und allen weiteren Informationen zu etwaigen Tests haben, die eine Bewertung des Testergebnisses ermöglichen, und dass die Aussagekraft von Tests so klar ist, dass Hersteller und Verwender von ADM-Systemen einerseits, Aufsichtsbehörden und gegebenenfalls auch Betroffene andererseits die Möglichkeit haben, Schlüsse auf Pflichten beim Einsatz des Systems zu ziehen. Diese Fragen müssen derzeit als offen bezeichnet werden.

Angesichts der offenen Fragen lässt sich insoweit insbesondere Forschungsbedarf feststellen. Gegenstand der Forschung sollten dabei nicht zuletzt die Qualität von Testverfahren sowie die Möglichkeit des Zugangs zu Informationen über Testergebnisse und deren Bewertung sein.

Angesichts der erheblichen Unklarheit ist anzunehmen, dass als Ergebnis der Forschung ein Gesetzgebungsbedarf festgestellt wird, der insbesondere der Konkretisierung der Pflichten von Herstellern und Nutzern von ADM-Systemen zu deren Überwachung einschließlich der Beachtung von Testergebnissen dienen sollte.

7.3.3 Durchsetzung der Durchführung von Tests

7.3.3.1 Notwendigkeit der rechtlichen Regelung

Ein wesentlicher Aspekt ist die Sicherung der Durchführung von Tests. Hersteller und Verwender von ADM-Systemen haben regelmäßig ein Interesse an der Durchführung von Tests der Systeme zur Qualitätssicherung und zur Vermeidung fehlerhafter algorithmischer Entscheidungen. Wie im Rahmen der Untersuchung des Begriffs des Fehlers und der systematischen Unterscheidung möglicher Fehler einer algorithmischen Entscheidung dargestellt, wird sich diese aber nicht notwendigerweise auf den Schutz des Betroffenen und dessen Persönlichkeitsrechte erstrecken.

Damit kann auf das Eigeninteresse von Herstellern und Verwendern von ADM-Systemen zur Durchführung von Tests zum Schutz der Persönlichkeitsrechte der Betroffenen nicht gesetzt werden.

Daher ist die Durchführung von Tests rechtlich abzusichern, damit gewährleistet ist, dass Tests zum Schutz der Betroffenen auch tatsächlich durchgeführt werden.

Die Durchsetzung des Tests von ADM-Systemen kann durch ganz unterschiedliche rechtliche Instrumente erreicht werden.

7.3.3.2 Tests als Zulassungsvoraussetzungen

Die Durchführung von Tests kann durch Gesetz als Zulassungsvoraussetzung für den Vertrieb oder den Einsatz von ADM-Systemen ausgestaltet werden. Derartige Regelungen sind im Bereich von Maschinen (etwa Kfz und Medizinprodukten) und Anlagen weit verbreitet. Dies geht regelmäßig mit einer intensiven rechtlichen Regelung des betreffenden Bereichs einher.

7.3.3.3 Zertifizierung und Prüfung

Die Durchführung von Tests hat große Bedeutung im Bereich von Zertifizierungen oder technischen Prüfungen von Maschinen und Anlagen. Zertifizierungen können, als Instrumente der Selbstregulierung, freiwillig erfolgen oder auch gesetzlich angeordnet sein, wie etwa im bekannten Beispiel der Kfz-Hauptuntersuchung. Zertifizierungen[295] dürften ein vielversprechendes Instrument zur Qualitätssicherung von ADM-Systemen sein. Freilich besteht noch Forschungsbedarf zur Frage, ob und gegebenenfalls wie eine Zertifizierung rechtlich erzwungen werden sollte oder rechtliche Anreize geschaffen werden sollten, wie es die DSGVO in Bezug auf die Zertifizierung für Auftragsverarbeiter tut. Soweit eine Pflicht zur Zertifizierung oder Prüfung angeordnet werden soll, ist eine gesetzliche Grundlage erforderlich.

7.3.3.4 Tests als Aufsichtsmaßnahme

Tests können von Aufsichtspersonen im Rahmen ihrer Aufsichtstätigkeit durchgeführt oder angeordnet werden. Dies setzt die – bisher nicht durchgehend gegebene – Existenz zuständiger Aufsichtsbehörden und entsprechender Aufsichtspersonen voraus.

7.3.3.5 Haftung als Anreiz zur Durchführung von Tests

Ein wesentlicher Anreiz zur Durchführung von Tests kann durch Haftungsregelungen geschaffen werden. Wie dargestellt, besteht eine Haftungsregelung dem Grundsatz nach bereits in Form des § 823 BGB. Jedoch könnte eine spezielle Haftungsnorm weitaus direkter und dosierter Anreize setzen. So könnte beispielsweise durch Gesetz eine Haftung der Produzenten durch fehlerhaft programmierte Entscheidungen geschaffen werden, bei der die Haftung wesentlich von der Durchführung eines geeigneten Tests abhängt.[296]

7.3.3.6 Beweislastregeln als Anreiz zur Durchführung von Tests

Ähnliche Steuerungseffekte wie Haftungsregeln können Beweislastregeln entfalten, die im Rahmen bestehender Haftungsregeln greifen. Das schärfste Instrument ist insoweit die Beweislastumkehr, die etwa dann reifen könnte, wenn kein ordnungsgemäßer Test durchgeführt wird. Gegenüber der in der Literatur vorgeschlagenen Beweislastumkehr hat ein solche Besonderheit eine andere Wirkung, indem sie gezielt auf die Durchführung von Tests drängt und damit für den Verwender von ADM-Systemen wesentlich geringere Haftungsrisiken schafft als eine allgemeine Beweislastumkehr.

7.3.3.7 Zivilrechtliche Ansprüche auf die Durchführung von Tests

Ein weiteres, überaus mächtiges Mittel zur Durchsetzung von Tests könnten zivilrechtliche Ansprüche auf die Durchführung oder Duldung von Tests darstellen. Ob und in welchem Umfang sich zivilrechtliche Ansprüche auf die Durchführung oder Duldung von Tests aus dem geltenden Recht ableiten lassen, ist unklar. Jedenfalls besteht keine gesetzliche Regelung, die Privaten einen Anspruch auf Durchführung eines Tests von ADM-Systemen gibt. Daher ist zu erörtern, ob derartige Rechte eingeführt werden sollten.

7.3.3.7.1 Vorteile und Risiken zivilrechtlicher Ansprüche auf Testdurchführung

[295] Zu Recht wird das Instrument der Zertifizierung in der Literatur empfohlen (dazu oben 7.1.1.10).
[296] In eine zumindest ähnliche Richtung geht der Vorschlag von Wischmeyer, AöR 2018, 1, 62, der Haftungserleichterungen bei regelmäßiger Durchführung von Audits erwägt.

Die Vorteile eines zivilrechtlichen Anspruchs auf Durchführung eines Tests sind offensichtlich: Die Berechtigung füllt eine Lücke des geltenden Rechts, das die Durchführung von Tests bisher nicht sichert. Da die Durchführung von Tests die Möglichkeiten zur Ermittlung von Fehlern maschineller Entscheidungen, insbesondere Diskriminierung, entscheidend verbessert, ist sie ein entscheidendes Mittel des Rechtsschutzes.

Der besondere Vorteil eines zivilrechtlichen Anspruchs auf Durchführung eines Tests von Entscheidungssystemen liegt darin, dass die von einer algorithmischen Entscheidung Betroffenen ihren Rechtsschutz selbst durchsetzen können und nicht etwa auf das Handeln Dritter, etwa einer Aufsichtsbehörde, angewiesen sind.

Die Einführung eines solchen zivilrechtlichen Anspruchs hätte freilich auch Nachteile. So ist die Wahrnehmung des Rechts auf Testen für beide Seiten aufwendig, selbst wenn man einmal annimmt, dass in Zukunft Testverfahren hinreichender Qualität mit angemessenem Aufwand möglich sind.

Ebenso müssen die Rechte des Produzenten und Verwenders von ADM-Systemen gewahrt sein, etwa im Hinblick auf Geschäftsgeheimnisse. Zwar führt die Durchführung eines Tests typischerweise nicht zur Offenlegung von Geschäftsgeheimnissen. Es besteht aber vermutlich ein Risiko: Jedenfalls durch „exzessives“ Testen wird man einen Einblick in die Struktur des Programms erhalten.

Es könnten auch durch unsorgfältige Tests missverständliche Testergebnisse produziert werden, deren Veröffentlichung für den Hersteller oder Verwender des ADM-Systems problematische Reputationsschäden hervorrufen kann. So kann man sich leicht vorstellen, dass durch einen Test eine Diskriminierung festgestellt wird, die sich bei einer Nachprüfung mit einem anderen Test als nicht haltbar erweist.

Es erscheint daher notwendig, ein etwaiges Recht zum Testen zu begrenzen, um die Rechte der Verwender von ADM-Systemen zu wahren.

7.3.3.7.2 Ausgestaltung eines Anspruchs auf Testdurchführung

Die konkrete Ausgestaltung eines Anspruchs auf Testdurchführung muss den widerstreitenden Interessen Rechnung tragen.

Anspruchsgegner sollte, soweit es um einen zivilrechtlichen Anspruch des Betroffenen geht, der Verwender des ADM-Systems sein.

Zunächst sollte der Anspruch nicht voraussetzungslos, sondern nur bei einem hinreichenden Verdacht auf einen Fehler, etwa eine Diskriminierung, bestehen. Dieser Verdacht wäre vom Anspruchsberechtigten substantiiert vorzutragen. Dies könnte im Fall der Überprüfung einer konkreten Kreditentscheidung etwa darin bestehen, dass Umstände vorgetragen werden, die für die Kreditwürdigkeit sprechen.

Eine entscheidende Weichenstellung eines Rechts auf Testdurchführung betrifft die Aktivlegitimation. Dies Recht kann als Individualanspruch dem von einer algorithmischen Entscheidung Betroffenen eingeräumt werden. Allerdings stellen sich dann die genannten Probleme wie Kosten des Testverfahrens oder etwaige Missbrauchsrisiken (Ausspähen) mit besonderer Schärfe.

Daher ist zu erwägen, ergänzend ein Recht zur Durchführung von Tests von ADM-Systemen bestimmten Organisationen, etwa Verbraucherschutzverbänden, einzuräumen oder gar auf

derartige Organisationen zu begrenzen. Letztlich dürfte, ohne dass dies im Rahmen dieser Studie abschließend beurteilt werden könnte, ein Recht auf Testdurchführung nur dann effizient sein, wenn es durch Organisationen wahrgenommen wird. Ein solches Recht sollte mit einem Unterlassungsanspruch der Organisation verbunden werden.

Bei der Testdurchführung sind sehr unterschiedliche Modelle denkbar. Eine einfache und zugleich weitreichende Maßnahme wäre etwa die Pflicht, eine Schnittstelle zum ADM-System bereitzustellen, über die z.B. nach eigenem Ermessen Tests gemacht werden können. Weniger weitreichend wäre es, solche Pflichten für Gerichtsverfahren aufgrund richterlicher Anordnung zu beschließen.

Die Regelung über ein individuelles Recht zur Durchführung von Tests wäre durch einen vorgelagerten Anspruch auf Auskunft für etwaige Testberechtigte zu ergänzen, damit ein derartiges Recht geltend gemacht werden kann.

7.3.3.8 Durchführung von Tests durch Beschwerde- oder Schiedsstellen

Denkbar erscheint es auch, statt eines individuellen Anspruchs auf Testdurchführung ein administriertes Verfahren zur Durchführung von Tests zu etablieren, wobei Tests auf Initiative von Betroffenen (Beschwerden) durch eine neutrale Stelle vorgenommen werden. Diese neutrale Stelle könnte eine staatliche Stelle, etwa eine Agentur für ADM-Verfahren, oder eine von der Wirtschaft zu betreibende Institution, ähnlich den Ombudsleuten im Bank- und Versicherungswesen, sein.

In diesem Fall könnten durch Einrichtung einer neutralen Instanz Missbrauchsrisiken eingedämmt und Vorteile durch eine Breitenwirkung der Tests erzielt werden. Die Unterhaltung und Finanzierung der Testverfahren könnte Herstellern und Verwendern von ADM-Systemen auferlegt werden, Missbräuche durch Antragsteller könnten durch Gebührenregelungen eingedämmt werden.

7.3.3.9 Zwischenergebnis

Der Überblick zeigt, dass die rechtlichen Möglichkeiten zur Durchsetzung von Tests von ADM-Systemen und zur Überprüfung algorithmischer Entscheidungen vielseitig und komplex sind. Eine umfassende Analyse dieses Aspekts kann im Rahmen dieser Studie nicht erfolgen.

7.3.4 Anforderungen an Testverfahren

Als ein Ergebnis der Studie hat sich ergeben, dass Tests von ADM-Systemen und die Überprüfung algorithmischer Entscheidungen erhebliche rechtliche Bedeutung haben. Voraussetzung hierfür ist aber, dass die Testverfahren als zuverlässig und die daraus erzielten Testergebnisse als richtig angesehen werden können.

Bei den damit angesprochenen qualitativen Anforderungen an Testverfahren sind zwei Aspekte zu unterscheiden: zum einen die Qualität des Testverfahrens, insbesondere des Prüfstandards, als solches, das insbesondere zuverlässig sein muss und eine hinreichende Aussagekraft aufweisen muss, und zum anderen die Durchführung eines konkreten Tests, insbesondere die korrekte Anwendung des Testverfahrens und ein ordnungsgemäßes Verfahren im Übrigen.

In beiden Aspekten muss eine hinreichende Qualität gesichert sein, damit Rechtsfolgen ausgelöst werden können.

Für den rechtlichen Rahmen von Testverfahren für ADM-Systeme bedeutet dies, dass diese Qualität rechtlich abgesichert oder jedenfalls für die Zwecke der Normen, für die Tests von Bedeutung sind, die erforderliche Qualität von Testverfahren so präzise zu bestimmen ist, dass das Vorliegen einer hinreichenden Qualität des Verfahrens überprüft werden kann.

7.3.4.1 Qualität von Testverfahren und Prüfstandards

Die Möglichkeiten des Tests von ADM-Systemen wurden im technischen Teil der Studie mit Blick auf Verfahren mit Maschinellem Lernen untersucht. Als ein Zwischenergebnis der Analyse ergibt sich, dass insoweit noch erheblicher Entwicklungsbedarf besteht (siehe Kapitel 4.5).

Aus rechtlicher Sicht ist es nicht nur für diesen wichtigen Spezialfall, sondern für ADM-Verfahren generell von Bedeutung, die erforderliche Qualität rechtlich zu fundieren oder rechtlich abzusichern. Ansonsten wären Tests jeweils im konkreten Einzelfall, etwa durch eine Aufsichtsperson, oder in einem Gerichtsverfahren, zu bewerten.

Der erste Schritt zu diesem Ziel müssen eine genaue Ermittlung sowie eine umfassende Beschreibung und Untersuchung der gegenwärtig vorhandenen Testverfahren für ADM-Systeme sein, die, soweit ersichtlich, bisher nicht vorliegen.

Darauf aufbauend sollte die notwendige Qualität von Testverfahren bestimmt werden. Dies könnte eine Aufgabe einer staatlichen Agentur für ADM-Systeme sein (siehe dazu unten 8.2.2), die derartige Prozesse orchestrieren und, soweit sie aus der Praxis angestoßen werden, daran teilnehmen könnte. Dies betrifft etwa die Mitarbeit in Standardisierungsgremien etc.

Auf dieser Grundlage sollte dann geprüft werden, welche legislativen Maßnahmen erfolgen können, um klare rechtliche Rahmenbedingungen hinsichtlich der Qualität an Testverfahren zu stellen.

7.3.4.2 Testverfahren

Die ordnungsgemäße Durchführung von Tests ist eine entscheidende Voraussetzung für Vertrauen in Testergebnisse, ebenso in die Möglichkeit, durch die Durchführung des Tests eine rechtliche Verpflichtung zu erfüllen.

Kernelemente eines ordnungsgemäßen Verfahrens technischer Überprüfung sind stets die Kompetenz und die Unabhängigkeit der Stelle, die den Test durchführt, sowie eine fachlich richtige Handhabung des Testverfahrens als solches. Hinzu kommen weitere verfahrensabhängige Anforderungen, wie etwa die Anhörung des Verwenders des Testobjekts sowie eine Widerspruchsmöglichkeit gegen Testergebnisse, die einer Regelung bedürfen.

Die Durchführung von Tests technischer Verfahren ist nicht durchgängig rechtlich geregelt. Je mehr einem Test rechtliche Bedeutung beigemessen wird, desto stärker ist freilich die Notwendigkeit, auch den Test zum Gegenstand rechtlicher Regelungen zu machen.

Der rechtliche Regelungsteppich für Tests technischer Systeme ist dicht und folgt einem Gewebe aus Vorgaben allgemeiner Regeln über Konformitätsbewertungen einerseits und bereichsspezifischen Regeln andererseits.

Zu den allgemeinen Regeln gehört etwa die Verordnung 765/2008[297] zur Akkreditierung von Konformitätsbewertungsstellen, deren Vorgaben in Deutschland durch das Akkreditierungsstellengesetz[298] umgesetzt wurden und das ein Akkreditierungsverfahren für Konformitätsbewertungsstellen vorsieht. Die VO 765/2008 wurde in Ergänzung des europäischen Beschlusses über einen gemeinsamen Rechtsrahmen für die Vermarktung von Produkten[299] erlassen und schließt insbesondere an die europäische Produktsicherheitsrichtlinie an.

Die VO 765/2008 und das AkkStelleG schließen andererseits Regeln über Konformitätsbewertungsstellen nicht aus. So stellt § 1 Abs. 2 AkkStelleG ausdrücklich fest, dass die aufgrund anderer Rechtsvorschriften bestehende Zuständigkeit zur Zulassung von Konformitätsbewertungsstellen unberührt bleibt.

Von Interesse ist, dass die Datenschutz-Grundverordnung in ihrem Art. 43 für den Bereich der datenschutzrechtlichen Zertifizierung auf die VO 765/2008 verweist. Entsprechend bezieht sich das neue Bundesdatenschutzgesetz in seiner Regelung zur Akkreditierung (§ 38) auf das AkkStelleG.

Die DSGVO liefert ein wichtiges Beispiel für verfahrensmäßige Anforderungen an Testverfahren im Zusammenhang mit dem Schutz von Persönlichkeitsrechten. In ihrem Art. 42 nennt sie wesentliche Grundlagen eines Zertifizierungsverfahrens:

- Transparenz des Zertifizierungsverfahrens (Art. 42 Abs. 3 DSGVO)
- Zertifizierung durch festgelegte Stellen (Art. 42 Abs. 5 DSGVO)
- Unabhängigkeit und Kompetenz der Zertifizierungsstellen (Art. 43 Abs. 1)
- Zertifizierung nach festgelegten Kriterien (Art. 42 Abs. 5 DSGVO)

Das Regelungsschema des Produktsicherheitsrechts dürfte auf ADM-Systeme zur Erzeugung algorithmischer Entscheidungen, wenn überhaupt, allenfalls in seltensten Einzelfällen anwendbar sein. Jedenfalls im Bereich des Scorings, des *dynamic pricing* oder der Berechnung von Versicherungsraten ist es nicht anwendbar.

Insgesamt ist überaus unklar, ob und in welchen Bereichen algorithmische Entscheidungen festgelegten Anforderungen an das Verfahren über die etwaige Überprüfung algorithmischer Entscheidungen oder Tests von ADM-Systemen unterliegen. Jedenfalls besteht insofern keine einheitliche Regelung.

[297] Verordnung (EG) Nr. 765/2008 des Europäischen Parlaments und des Rates vom 9. Juli 2008 über die Vorschriften für die Akkreditierung und Marktüberwachung im Zusammenhang mit der Vermarktung von Produkten und zur Aufhebung der Verordnung (EWG) Nr. 339/93 des Rates (Text von Bedeutung für den EWR), ABl. L 2008 Nr. 218/30.
[298] Gesetz über die Akkreditierungsstelle (Akkreditierungsstellengesetz – AkkStelleG) vom 31.07.2009, BGBl. I S. 2625.
[299] Beschluss Nr. 768/2008/EG des Europäischen Parlaments und des Rates vom 9. Juli 2008 über einen gemeinsamen Rechtsrahmen für die Vermarktung von Produkten.

Angesichts der Bedeutung von Tests ergeben sich damit zwei Aufgabenbereiche: Zum einen gibt es Forschungs- und Entwicklungsbedarf in Bezug auf geeignete Testverfahren und zum anderen besteht Regelungsbedarf hinsichtlich der Anforderungen an das Testverfahren.

7.4 Transparenz und Information

Ein wesentlicher Gesichtspunkt für den rechtlichen Schutz gegenüber algorithmischen Entscheidungen sind Transparenz und Information des Betroffenen.

Transparenz und Information über ADM-Systeme sind nicht nur von Bedeutung, wenn eine Entscheidung durch eine Maschine ohne menschliches Zutun getroffen wird, sondern auch dann, wenn ein ADM-System zur Entscheidungsvorbereitung genutzt wird. Dies ist etwa der Fall, wenn die Entscheidung über eine Kreditvergabe durch einen menschlichen Sachbearbeiter getroffen wird, dieser sich aber auf ein durch ein ADM-System ermitteltes Scoring des Antragstellers stützt.

7.4.1 Kennzeichnungspflicht für ADM-Systeme

Die Transparenz kann sich auf die Tatsache des Einsatzes eines ADM-Systems beziehen. Insoweit wird in der Literatur eine Kennzeichnungspflicht vorgeschlagen. Dieser Vorschlag ist zumindest erwägenswert, da weitere Rechte und Schutzmechanismen, die vom Betroffenen oder Dritten geltend zu machen sind, nur greifen können, wenn die Tatsache überhaupt bekannt ist.

Eine allgemeine Kennzeichnungspflicht für den Einsatz von ADM-Systemen, die nach geltendem Recht nicht besteht, wäre durch spezielle Gesetze einzuführen. Insoweit ist eine Reihe von Aspekten zu berücksichtigen.

7.4.1.1 Anwendungsbereich

Voraussetzung einer gesetzlichen Kennzeichnungspflicht ist, dass die kennzeichnungspflichtigen Systeme deutlich bezeichnet werden. Bei dieser schwierigen Aufgabe wäre zugleich zu prüfen, ob eine flächendeckende Kennzeichnung sachgerecht ist. Gegen eine umfassende Kennzeichnungspflicht spricht, dass eine Kennzeichnung nicht geboten ist, soweit schutzwürdige Interessen Dritter nicht betroffen sind. Dabei ist auch zu beachten, dass eine Kennzeichnung faktisch leerläuft, wenn sie nicht differenziert. Um ein Beispiel zu nennen: Der Hinweis, dass ein Text unter Verwendung elektronischer Daten erzeugt wurde, wäre in der digitalen Gesellschaft dysfunktional. Abzuwägen ist dann weiter, ob die Kennzeichnung in jedem Fall zumutbar ist. Der Einsatz innovativer Technik ist in vielen Fällen ein Wettbewerbsvorteil, der grundsätzlich auch schützenswert ist.

Kennzeichnungspflichten sind nur sinnvoll, wenn die damit transportierte Information relevant ist. Kennzeichnungspflichten sollten daher bestehen, wenn damit auch andere Rechte, etwa auf Überprüfung, einhergehen. In diesen Fällen handelt es sich um eine notwendige Ergänzung des Rechtsschutzes. Ob darüber hinaus Kennzeichnungspflichten sinnvoll sind, ist fraglich und wäre genauer zu prüfen.

7.4.1.2 Inhalt und Bekanntmachung der Kennzeichnung

Soweit in der Literatur eine Kennzeichnungspflicht für ADM-Systeme vorgeschlagen wird, fehlt eine nähere Beschreibung des Inhalts der Pflicht.

Derzeit ist daher unklar, welche Informationen mitzuteilen sind. Als Minimalinhalt kann die Tatsache, dass ein ADM-System eingesetzt wird, angenommen werden. Es ist aber offensichtlich, dass weitere Informationen über die Art des Systems, den Einsatzbereich etc. sinnvoll sein können. Dies bedarf einer näheren Untersuchung.

Weiter ist unklar, wie die Kennzeichnung bekanntzumachen ist. Denkbar, aber wohl eher nicht zweckmäßig ist ein Register für ADM-Systeme. Überzeugender wäre wohl eine Mitteilung an Personen, die vom ADM-System betroffen sein können, oder an eine Aufsichtsbehörde.

7.4.2 Informationspflichten

Der Pflicht zur Kennzeichnung ähnlich oder je nach Ausgestaltung gleich sind Informationspflichten beim Einsatz von ADM-Systemen. Als Informationspflicht werden hier, in Abgrenzung zu Kennzeichnungspflichten, solche Pflichten bezeichnet, die eine an einen bestimmten Empfänger, etwa einen Vertragspartner, zu richtende Information anordnen.

Soweit algorithmische Entscheidungen gegenüber Verbrauchern eingesetzt werden, sollte statt einer allgemeinen Kennzeichnung eine konkrete Information im Zusammenhang mit der algorithmischen Entscheidung erfolgen.[300] Dabei wäre es sinnvoll, mit der Information über den Einsatz eines ADM-Systems auch Hinweise zu etwaigen Aufsichtsstellen etc. zu geben.

Soweit man die Einführung von Informationspflichten erwägt, ist deren Umfang zu klären. Besonders weitgehend ist hier der Vorschlag von Martini, eine Begründung algorithmischer Entscheidungen zu verlangen.[301] Im geltenden Recht enthalten die Art. 13 und 14 DSGVO bereits Informationspflichten bei automatisierten Entscheidungen (Art. 13 Abs. 2 lit. f und Art. 14 Abs. 2 lit. g DSGVO). Allerdings ist unklar, ob die Informationspflichten nur bei ausschließlicher automatisierter Verarbeitung (siehe dazu oben 5.5.6) und Profiling oder darüber hinaus bei jeglichem Einsatz von ADM-Systemen greifen. In inhaltlicher Weise sind die Informationspflichten recht umfassend; zu informieren ist nicht nur über das „Ob“ einer automatisierten Entscheidungsfindung, sondern auch über wesentliche Aspekte des verwendeten Algorithmus. Daher ist im Lichte der Auslegung der Art. 13 und 14 DSGVO zu untersuchen, ob insoweit Lücken bestehen. Insbesondere ist zu erwägen, eine entsprechende Informationspflicht auch in Fällen der Entscheidungsvorbereitung durch ADM-Systeme einzuführen.[302] Eine Beschränkung der Informationspflicht auf Fälle des Profiling, wie sie für Art. 22 DSGVO angenommen werden kann, sollte jedenfalls nicht erfolgen, da problematische Ungleichgewichte durch den Einsatz von ADM-Systemen auch außerhalb von Profiling entstehen können.

Hinsichtlich der Informations- und Kennzeichnungspflichten sind insbesondere der Anwendungsbereich zu berücksichtigen. Bei einer Ausdehnung der Informationspflicht auf die Entscheidungsvorbereitung durch ADM-Systeme ergeben sich notwendigerweise schwierige Abgrenzungsfragen. Eine uferlose Ausdehnung der Informationspflicht ist

[300] So auch Martini, JZ 2017, 1017, 1020.
[301] Martini, JZ 2017, 1017, 1020; siehe dazu auch oben Abschnitt 7.1.1.3.
[302] Ebenso Wischmeyer, AöR 2018, 1, 23.

unbedingt zu vermeiden, schon weil die Information nicht mehr wirken kann, wenn sie unterschiedslos bei jeder Art von Entscheidung erteilt wird. Eine Informationspflicht sollte daher nur angeordnet werden, soweit ein entsprechender Schutzbedarf des Verbrauchers besteht. In den in dieser Studie genannten Fallgruppen des Kreditscorings und individueller Versicherungstarife ist ein solcher Schutzbedarf zu bejahen.

7.4.3 Zwischenfazit

Als Ergebnis zeigt sich damit, dass Kennzeichnungs- und Informationspflichten bei ADM-Systemen ein wichtiges Element der Reglementierung sind.

Eine gesetzliche Kennzeichnungspflicht für ADM-Systeme kann jedenfalls in Teilbereichen sinnvoll sein. Insbesondere erscheinen gesetzliche Informationspflichten, soweit sie nicht ohnehin bestehen, gegenüber Verbrauchern sinnvoll.

Allerdings stehen auch diese im Kontext mit anderen Maßnahmen und sollten daher im Zusammenhang mit diesen zunächst weiter erforscht und sodann in einem Gesetzgebungsakt umgesetzt werden.

7.5 Zusammenfassung der Ergebnisse anhand der Gefährdungsszenarien

Entsprechend der Aufgabenstellung des SVR liegen der Studie drei Gefährdungsszenarien zugrunde.[303] Nachfolgend werden die Ergebnisse der Kapitel 4, 5 und 7 auf diese Gefährdungsszenarien bezogen.

7.5.1 Gefährdungsszenario 1: Inhaltlich unrichtige Entscheidung

Inhaltlich unrichtige Entscheidungen durch ADM-Systeme liegen vor, wenn auf Basis der zur Verfügung stehenden Daten die falsche Entscheidung getroffen wurde. Es geht hier um den eigentlichen Entscheidungsausgang (z.B. ein Kredit wird einem faktisch kreditwürdigen Antragsteller verweigert).

Kapitel 4.3 und 5.3 erklären und illustrieren, dass inhaltlich falsche Entscheidungen in verschiedene Fehlertypen unterteilt werden können und ein Regulierungsregime daran anknüpfen kann. Das Fehlerverhalten eines ADM-Systems kann also anhand von Testdaten genau untersucht wurden, sofern diese Testdaten repräsentativ für die praktische Anwendung des Systems sind. Kapitel 4.4 erläutert technische Verfahren, um durch Tests und Audits ADM-Systeme auf deren inhaltliche Richtigkeit zu überprüfen.

ADM-Systeme werden auf Testdaten praktisch niemals fehlerfrei entscheiden. Rechtlich stellen sich nun, wie in Kapitel 5 dargestellt, mehrere Fragen: Ist der Einsatz eines ADM-Systems (ohne Überprüfung durch einen Menschen im Einzelfall) überhaupt zulässig? Unter Umständen könnte Art. 22 DSGVO dies einschränken, wie in Kapitel 5.5 erörtert. Die Reichweite der Norm ist jedoch beschränkt und in der Praxis zumindest der hessischen Datenschutzaufsicht spielt sie offenbar nur eine geringe Rolle.

[303] Siehe oben Kapitel 2.2.

Falls der ADM-Einsatz zulässig ist, können sich aus verschiedenen Gesichtspunkten Anforderungen an die Gestaltung des Entscheidungsprozesses ergeben. Dabei ist an eine mögliche strukturelle Überlegenheit des Entscheiders zu denken, wie sie in Kapitel 5.1 erörtert wird. Ob dieser und andere Gesichtspunkte eine neue Form der Algorithmenregulierung nahelegen – und welche Möglichkeiten hierfür bestehen –, diskutieren wir im vorliegenden Kapitel 7.

Aber auch bestehende gesetzliche Regelungen, wie sie etwa die Diskriminierung betreffen, stellen Anforderungen an das ADM-System. Dies führt uns zu Gefährdungsszenario 2.

7.5.2 Gefährdungsszenario 2: Diskriminierender Algorithmus

Das Beispiel der Diskriminierung macht deutlich, dass es keine einfache Lösung gibt, um die Richtigkeit von Entscheidungen sicherzustellen. So können beim Einsatz Maschinellen Lernens, wie wir in Kapitel 4.2 darstellten, schon die Trainingsdaten unausgewogen sein, was sich dann im Regelfall auch im gelernten Modell niederschlagen wird. Es existieren technische Ansätze zur Vermeidung dieses Effekts, die jedoch nur teilweise reif bzw. geeignet sind, um regulatorisch durch Gesetzgebung und Prozessgestaltung in der Aufsicht aufgegriffen zu werden. Ein Verzicht auf die Verwendung geschützter Attribute wie Geschlecht oder ethnische Herkunft kann die unmittelbare Benachteiligung (vgl. zu diesem Begriff Kapitel 5.4.1) der entsprechenden Gruppen verhindern. Aufgrund der Zusammenhänge mit anderen, nicht geschützten Attributen kann eine mittelbare Benachteiligung (vgl. Kapitel 5.4.2) aber nicht ausgeschlossen werden.

Um, soweit möglich, auf unerwünschte Benachteiligungen prüfen zu können, müssen einerseits operationalisierbare Kriterien vorliegen, die eine Implementierung der Anforderungen seitens der Verwender und eine Prüfung durch Behörden praktisch ermöglichen. Allgemein anerkannte, operationalisierbare Kriterien sind in der bisherigen Rechtsprechung und Literatur aber noch nicht ersichtlich. Eine Bestimmung dieser Kriterien kann durch die laufende Forschung zum Thema *Fair Machine Learning* erfolgen, beinhaltet aber notwendigerweise konkrete inhaltliche Ausgestaltungen des Gleichbehandlungsbegriffs in relevanten Rechts- und Sachgebieten. Wir gehen auf diese Problematik und den weiteren Forschungsbedarf oben in Kapitel 5.4.3 ein.

Andererseits bedarf es auch einer technischen Umsetzung. Bisher werden oft einfache, transparente Modelle, insbesondere basierend auf der logistischen Regression, eingesetzt (vgl. Kapitel 4.1.2). Hier können die Modelle noch mit vertretbarem Aufwand durch Experten analysiert werden. Jedoch stellt sich das Problem möglicher Diskriminierung allgemeiner. Ansätze liegen im Testen bzw. der Auditierung, für die aber ein Zugriff auf das ADM-System sowie geeignete Testdaten benötigt werden. Ob eine Prüfung bzw. ein Test ex ante vorgesehen werden sollte und wie ein rechtlicher Rahmen für Tests aussehen könnte, diskutieren wir in Kapitel 7.3.

Für die Gefährdungsszenarien 1 und 2 gleichermaßen ist es von Bedeutung, wie ein Fehler einer algorithmischen Entscheidung, etwa eine Diskriminierung, festgestellt werden kann. Als Mittel können dazu insbesondere eine Code-Analyse sowie ein Test des ADM-Systems eingesetzt werden (vgl. Kapitel 7.2.2) Als besonders vielversprechend wird in der Studie das Testen erkannt, das auch bei selbstlernenden Systemen angewendet werden kann (vgl. Kapitel 7.2.2.3). Insbesondere können Tests in den Gefährdungsszenarien 1 und 2 eingesetzt werden.

Soweit demnach dem Testen von ADM-Systemen eine wesentliche Bedeutung für die rechtliche Regelung von ADM-Systemen zukommen soll, ist es erforderlich, einen rechtlichen Rahmen für das Testen zu schaffen, der sowohl die Qualität des Testverfahrens als auch die Durchführung von Tests sicherstellt und die rechtliche Bedeutung von Tests festlegt (vgl. Kapitel 7.3). Insoweit gelangt die Studie zu dem Ergebnis, dass in allen Aspekten noch erheblicher Bedarf an Forschung und gesetzlicher Regelung besteht, um für die Gefährdungsszenarien 1 und 2 eine effektive Bekämpfung fehlerhafter Entscheidungen und diskriminierender Algorithmen sicherzustellen.

7.5.3 Gefährdungsszenario 3: Intransparent personalisierender Algorithmus

Unsere Literaturanalyse und Befragungen deuten an, dass das Problem der Preisdiskriminierung zwar viel diskutiert wird, das Ausmaß und die angewandten Methoden aber weitgehend unbekannt sind. Gerichtsentscheidungen gibt es hier noch nicht, auch nicht wenn die internationale Dimension mit einbezogen wird. Die bestehende Diskussion bleibt daher weitgehend spekulativ. So verwendet die Artikel-29-Datenschutzgruppe Preisdiskriminierung zwischen Apple- und PC-Benutzern (ausgehend von der Idee, dass sich Erstere größere Ausgaben leisten können) als (problematisches) Beispiel für eine Verletzung des Grundsatzes der Eingrenzung des Verwendungszwecks, ohne aber auf konkrete Beispiele zu verweisen. Ob es hier zu Problemgestaltungen kommt, die von denen im Kreditscoring und Datenschutzrecht unterschiedlich sind und einer gesonderten Regelung bedürfen, wird von diesen spezifischen Fragen der Ausgestaltung der Algorithmen abhängig sein. Es gibt es in verschiedenen Branchen zahlreiche Fälle von legitimer Preisdifferenzierung (Risikogruppen, Angebotsauslastung etc.) und ob der Einsatz von Algorithmen hier neuen Handlungsbedarf erzeugt – über das in dieser Studie Verlangte hinaus –, würde eine sehr viel detailliertere empirische Analyse der Methoden, Verbreitung und Auswirkungen auf den Verbraucher verlangen. So ist es bereits umstritten, ob sich Preisdifferenzierung generell positiv oder negativ auf Verbraucher auswirkt. Auch aus diesem Grund ist die Debatte zur algorithmischen Preisdifferenzierung auch international bislang schwerpunktmäßig im Wettbewerbs- und Kartellrecht geführt worden.

Für unsere engere Fragestellung indes sind die Ausführungen zum Kreditscoring und Datenschutz analog anwendbar. Insbesondere bedeutet dies, falls wirklich ein Verbraucher als wohlhabender klassifiziert wird und dann einen einfachen Artikel mit höheren Preisen angezeigt bekommt, ohne dass dabei ein Risiko-, Nachfrage- oder Vorratsaspekt gegeben wäre, dies per Sock Puppet Audit im laufenden System möglicherweise gut prüfbar ist. Dazu muss man „nur“ mehr und weniger wohlhabende Benutzer „simulieren“.[304] Hier gilt natürlich, dass von einem als Profiling erscheinenden Verhalten nicht automatisch auf den Rechtsverstoß geschlossen werden kann. Ist eine solche Prüfung nicht möglich (etwa weil eine Benutzerhistorie nicht retroaktiv erschaffen werden kann), dann kann eine nicht-datenschutzrechtliche Prüfung nur durch ein Audit des Modells selbst (z.B. anhand von Unternehmensdaten) im Unternehmen stattfinden, wobei hier unter Umständen Geschäftsgeheimnisse eine Rolle spielen können.

[304] Eine Studie, die sowohl die Möglichkeit des Testens als auch empirisch die Prävalenz und Methode erforscht, ist unter anderem die von Aniko Hannak, Gary Soeller, David Lazer, Alan Mislove und Christo Wilson. „Measuring price discrimination and steering on e-commerce web sites.“ In Proceedings of the 2014 conference on internet measurement conference, S. 305-318. ACM, 2014.

Derzeit besteht bereits erhebliche Unklarheit über die rechtlichen Grenzen des Einsatzes von ADM-Systemen zur Preisdiskriminierung, die nicht Gegenstand der Studie sind. Im Hinblick auf das damit zusammenhängende allgemeine Problem der intransparent personalisierenden Algorithmen werden Kennzeichnungs- und Informationspflichten diskutiert (vgl. Kapitel 7.4). In der Studie werden gesetzliche Kennzeichnungspflichten für den Einsatz von ADM-Systemen als grundsätzlich geeignete Mittel zur Verbesserung der Transparenz angesehen, jedoch nur für bestimmte, konkrete Fallgruppen empfohlen: Diese sollten, über die Regelung der DSGVO hinaus, eingeführt werden, soweit Rechte auf Überprüfung algorithmischer Entscheidungen bestehen. Dies kann aber beispielsweise für Preisdiskriminierung derzeit nicht ohne weiteres als gegeben angesehen werden. Ähnliches gilt für Informationspflichten, die nur punktuell, bei besonderem Schutzbedarf etwa von Verbrauchern, eingeführt werden sollten (vgl. Kapitel 7.4.2).

Damit stellt die Studie auch zum im Gefährdungsszenario 3 angesprochenen Aspekt der Transparenz erheblichen Forschungsbedarf fest (vgl. Kapitel 7.4.3).

8 Handlungsempfehlungen

Die Studie zeigt, dass bei ADM-Systemen und algorithmischen Entscheidungen Handlungsbedarf auf mehreren Gebieten besteht. Die erforderlichen Maßnahmen werden nachfolgend für die Bereiche der Forschung, generell des Erkenntnisgewinns zu ADM (8.1), organisatorischer Maßnahmen (8.2) sowie legislativer Maßnahmen (8.3) beschrieben.

8.1 Forschung, Ausbildung und Standardisierung

8.1.1 Bedarf an interdisziplinärer Forschung zu maschinellen Entscheidungen

Die interdisziplinäre Rechtsinformatikforschung zeichnet sich durch eine enge Zusammenarbeit zwischen Informatik und Rechtswissenschaft aus, weil nur dadurch der Stand der Wissenschaft weiterentwickelt werden kann.

Maschinelle Entscheidungen sind ein derzeit sehr aktueller Teil dieser Kooperation. Das Szenario der Entscheidungsprozesse ändert sich wesentlich. Während bisher Fehler von Menschen und ihre Vermeidung im Vordergrund standen, sind nunmehr Fehler beim Einsatz von Algorithmen zu berücksichtigen. Zudem werden maschinelle Entscheidungen häufig als Blackbox angesehen. Dem können nur eine bessere Erklärbarkeit und Auditing/Zertifizierung der maschinellen Entscheidungen Abhilfe schaffen.

Es sind große Forschungsanstrengungen in interdisziplinären Programmen erforderlich, in denen Informatiker, Mathematiker und Juristen die Themen maschineller Entscheidungen wissenschaftlich erforschen: So existieren beispielsweise umfangreiche technische Literatur zur Fairness algorithmischer Entscheidungen sowie juristische Literatur und Rechtsprechung, die die Feststellung von Diskriminierung durch menschliche Entscheider betrifft; wie die jeweils verwendeten Konzepte zusammenhängen und welche juristischen Anforderungen an eine diskriminierungsfreie maschinelle Entscheidung bestehen, ist aber noch unklar. Durch gezielte Förderung (Forschungsprojekte, Stipendien, Kollegien, Workshops, Hackathons, Challenge-Datasets etc.) muss ein Brückenschlag zwischen den Disziplinen vollzogen werden, der das gegenseitige Verständnis stärkt, eine gemeinsame Sprache für einen produktiven Austausch ermöglicht und so die Grundlage für eine ergebnisorientierte Forschung schafft. Ein naheliegendes Forschungsziel wäre beispielsweise eine rigorose empirische Untersuchung der praktischen Anwendbarkeit von quantitativen Fairnessmetriken und Fair-Machine-Learning-Methoden in konkreten Sachgebieten, in denen bereits jetzt ADM-Verfahren eingesetzt werden.

Solche Anforderungen könnten sich auch von denen an menschliche Entscheidungen unterscheiden, da eine genauere Analyse möglich wird, die gegebenenfalls zu schnell zu einer Einschätzung eines ADM-Systems als diskriminierend führt. Zu klären sind zudem Transparenzanforderungen an maschinelle Entscheidungen, Beweisverfahren bei möglicherweise diskriminierenden Algorithmen, die Relevanz von quantitativen Gleichbehandlungsbegriffen und deren praktische Nutzung, das Audit von bereits operablen ADM-Systemen etc. Im Ergebnis soll dies einen Katalog der wichtigsten Kriterien zur

Bewertung von ADM ergeben. Dieses Gutachten erläutert spezifische Aspekte dieses Forschungsbedarfs in den Kapiteln 5.4.3 und 7.3.

Diese Forschungsergebnisse sollen dann in der weiteren Erforschung zum Rechtsschutz gegenüber ADM-Verfahren münden. Aus Sicht der Rechtsinformatikexperten der Gesellschaft für Informatik ist eine umfangreiche Gap-Analyse des Rechtsrahmens erforderlich, um die Regelungslücken des bisherigen Datenschutzrechts und Diskriminierungsschutzes identifizieren zu können. Es ist zu prüfen, ob Verpflichtungen zur Transparenz, Auditing von Algorithmen, Beweiserleichterung bei Rechtsstreitigkeiten oder Beweislastumkehr hilfreich sind, um die Rechtsposition der Verbraucher zu verbessern, ohne eine ungebührliche Last für die betroffenen Unternehmen darzustellen oder teure bürokratische Strukturen schaffen zu müssen.

Die etablierten Konferenzen über AI und Recht – z.B. die *International Conference on Artificial Intelligence & Law* (ICAIL) sowie JURIX – sollten sich stärker dem Austausch der Forschungsergebnisse über den Einsatz von Algorithmen widmen. Dies kann durch die Unterstützung von Workshops gefördert werden. Hierbei sollen sowohl Algorithmen, Tests und Auditing-Verfahren als auch deren rechtliche Zulässigkeit jeweils interdisziplinär behandelt werden.

8.1.2 Verankerung in der Lehre und Ausbildung

Die Praxis des zunehmenden Einsatzes von ADM-Verfahren muss sich auch in der Lehre und Ausbildung an den Hochschulen niederschlagen. Besonders betroffen sind die Studiengänge der Informatik und Rechtswissenschaften. Bei Letzteren besteht der stärkere Handlungsbedarf, weil die Kenntnis von Algorithmen und deren rechtlicher Bewertung derzeit noch kaum gelehrt wird. Darüber hinaus gilt es, die Kompetenzen im Umgang mit digitalen Technologien und Daten in der Breite der Hochschulausbildung zu verankern – auch jenseits der Informatik und Rechtswissenschaften.[305]

8.1.2.1 Interdisziplinäre Programme an Hochschulen

Für die Ausbildung von Experten in der ADM wäre es notwendig, interdisziplinäre Programme an Hochschulen in Form von Fächerkombinationen und postgradualen Programmen einzurichten. Diese Programme sollen sich an Juristen, Wirtschaftswissenschaftler und Informatiker richten und sowohl den technischen Einsatz von ADM als auch die rechtlichen Fragen umfassen. Juristen sollten sich mit der Wechselwirkung zwischen Recht und angewandter Algorithmik und Machine Learning bereits im Studium als Schwerpunkt oder durch Seminare beschäftigen können. Studierende

[305] Außer den bereits genannten Disziplinen zeigt unsere Literaturanalyse, dass auch eine Integration mit den Wirtschaftswissenschaften verstärkt werden sollte. Entscheidungstheoretische Ansätze aus den Wirtschaftswissenschaften haben zumindest das Potenzial, informationstechnische, rechtliche und empirisch-wirtschaftliche Perspektiven zu verbinden. Wir haben diese in dieser Studie nur kurz anreißen können, doch typische Studien reichen von frühen Versuchen, diskriminierendes Verhalten von Juroren zu identifizieren (Finkelstein, Michael O., „The application of statistical decision theory to the jury discrimination cases." Harvard Law Review (1966), S. 338-376) bis zur gegenwärtigen Forschung im maschinellen Lernen (siehe z.B. Faisal Kamiran, Asim Karim und Xiangliang Zhang, „Decision theory for discrimination-aware classification." In *Data Mining (ICDM), 2012 IEEE 12th International Conference on*, S. 924-929. IEEE, 2012.) Dies ist insbesondere in Bereichen wie der Preisdiskriminierung hilfreich, in denen die juristische Einordnung des „Schadens" schwerfällt und das Problem eher die Auswirkung auf den Markt und den Wettbewerb als auf den einzelnen Verbraucher sein könnte.

der (Wirtschafts-) Informatik sollten die hier behandelten juristischen Aspekte der Anwendung von Algorithmen und regulierten datenbasierten Systemen als Nebenfach belegen oder ebenfalls in speziellen Seminaren erlernen können. Idealerweise sollte dies im Rahmen einer Zusammenarbeit zwischen den jeweiligen Fakultäten an den Hochschulen stattfinden und sollten die Teilnehmer schon im Studium dem Fachwissen, den Methoden und der Wissenschaftskultur beider Disziplinen ausgesetzt sein.

8.1.2.2 Übergreifende Kurse und Zertifikate

Zusätzlich wäre es wichtig, Aus- und Weiterbildungskurse für bereits qualifizierte Fachkräfte anzubieten, in denen die technischen Grundlagen sowie rechtlichen Rahmenbedingungen für ADM vermittelt werden. Auch der wirtschaftswissenschaftliche Hintergrund könnte dabei berücksichtigt werden. Die Attraktivität der Kurse kann durch Zertifikate und anerkannte Zusatzqualifikationen gesteigert werden. Die Kurse sollen sich inhaltlich sowohl an Juristen als auch an Informatiker und Ingenieure richten. Damit erreicht man nicht nur eine Steigerung des Problembewusstseins, sondern fördert implizit auch einen Diskurs zwischen beiden Disziplinen und erhöht die Vernetzung und Mobilität qualifizierten Personals.

8.1.2.3 Ausbildungsschwerpunkte für Juristen und Informatiker

In der Rechtsinformatik sollte Grundlagenwissen zu ADM als Teil der Ausbildung zum Einsatz von Technologien in der Rechtswissenschaft vermittelt werden. Zusätzlich wäre es hilfreich, auch Workshops zum praktischen Einsatz von ADM anzubieten; hier wäre ein Schwerpunkt auch auf die Einführung in die Programmierung von Algorithmen zu legen. Bei Informatikern sollten verstärkt die rechtlichen Grundlagen des Einsatzes von ADM, z.B. Datenschutz oder Gleichbehandlung, gelehrt werden.

8.1.2.4 Data Literacy in der wirtschafts- und sozialwissenschaftlichen Ausbildung fördern

Auch in anderen Disziplinen mit Schnittmengen zu den Rechtswissenschaften und der Informatik wie beispielsweise den Wirtschafts- und Sozialwissenschaften sollte die Vermittlung von Kompetenzen im Umgang mit ADM-Verfahren, die sie speisenden Daten und zugrundeliegenden Prozesse und Wirkungsweisen einen höheren Stellenwert bekommen. Eine Studie der Dalhousie University in Kanada fasst diese Fähigkeit unter dem Begriff „*Data Literacy*“ zusammen.[306] Dies beinhaltet die Kompetenzen, Daten erfassen, erkunden, managen, kuratieren, analysieren, visualisieren, interpretieren, kontextualisieren, beurteilen und anwenden zu können. *Data Literacy* wird als eine zentrale Kompetenz für die Digitalisierung und die globale Wissensgesellschaft angesehen.[307]

Die Fähigkeit, planvoll mit Daten umzugehen und sie im jeweiligen Kontext bewusst einsetzen und hinterfragen zu können, wird über viele Studienrichtungen – insbesondere für die Sozialwissenschaften inklusive der Rechts-, Wirtschafts-, Politik- und Verwaltungswissenschaften – hinweg immer wichtiger. Demnach ist *Data Literacy* eine zentrale Kompetenz für die Digitalisierung und die globale Wissensgesellschaft in allen Sektoren und Disziplinen. Angesichts der zunehmenden Menge und der Verfügbarkeit von Daten und der steigenden algorithmischen Durchdringung vieler Lebensbereiche stellt sich die Herausforderung, mit den Daten Wissen zu generieren, fundiert Entscheidungen treffen, aber auch ADM-Verfahren hinterfragen zu können.

[306] Ridsdale et al. 2015.
[307] Heidrich et al. 2018.

Wissen über Systeme und der kritische Umgang mit diesen müssen zudem in den Wirtschaftswissenschaften, speziell dem Management, gelehrt werden: einerseits weil in den Wirtschaftswissenschaften auch Entscheidungstheorie gelehrt wird, andererseits weil im Management die Begehrlichkeit, die Systeme zur Kosteneinsparung einzusetzen, vermutlich am größten ist. Gerade Einsätze zur Kostenersparnis können aber die problematischsten Verwendungen sein.[308]

8.1.3 Forschungsstrategie

8.1.3.1 Klassische Forschungsförderung

Im Rahmen dieser Arbeit wurden umfassend offene Fragen an der Schnittstelle zwischen Recht und Informatik aufgezeigt. Weiterführende Forschung ist erforderlich, um die rechtlichen Rahmenbedingungen klarer aufzuzeigen und die technische Machbarkeit sowie die zu erwartende Innovation umfassender zu verstehen.[309]

Es ist notwendig, dass dazu Forschungsprogramme in bestehenden Wissenschaftsförderstrukturen, z.B. durch das Bundesministerium für Bildung und Forschung (BMBF) oder die Deutsche Forschungsgemeinschaft (DFG), ins Leben gerufen werden, die sich dezidiert diesem Schnittstellenthema widmen. Man könnte diese zum Teil auch in den bereits existierenden Forschungsprogrammen über Künstliche Intelligenz oder Maschinelles Lernen integrieren. Das Entscheidende dabei wird es sein, dass der interdisziplinäre Charakter der Forschung gefördert wird und die Mittel nicht doch in weitgehend monodisziplinären Unternehmungen münden.

Vor dem Hintergrund, dass es sich um eine internationale Herausforderung handelt, muss auch geprüft werden, ob sich diese Thematik nicht auch in den großen Digitalisierungsinitiativen auf europäischer Ebene, z.B. „Horizon 2020“ oder ERC, niederschlagen sollte. Wenn die Forschungs- und Ausbildungsempfehlungen als Teil einer länderübergreifenden Initiative innerhalb Europas verstanden werden, kann dies zu großen Synergieeffekten führen und die Harmonisierung der europäischen Rechtsordnungen in Bezug auf die Herausforderungen der Digitalisierung fördern.

8.1.3.2 Koordinierte Auftragsforschung

Von europäischer bzw. deutscher Seite sollte – nach Konsultation der betroffenen Kreise – ein Forschungsplan (Roadmap) mit konkreter Zieldefinition über einen Zeitraum von mindestens drei Jahren ausgearbeitet werden. Zudem sollte ein umfangreiches Forschungsprogramm mit Auftragsforschung (Studien) zu Teilaspekten erstellt werden, das durch eine koordinierende Instanz bzw. ein Expertengremium gesteuert wird. Wesentlich ist die Einbindung von Gesetzgebern bzw. Ministerien als Bedarfsträger.

In regelmäßigen Konferenzen sollte der Abgleich der Ergebnisse bzw. die Fortentwicklung der Forschungsfragen erfolgen. Die Resultate sollten kontinuierlich publiziert und erfolgreiche Teilergebnisse von den Bedarfsträgern (probeweise) implementiert werden.

[308] So etwa Harkens 2018. Als Beispiel zur cost-benefit Analyse von algorithmischer Entscheidung siehe etwa Corbett-Davies, Sam, et al. "Algorithmic decision making and the cost of fairness." Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2017.
[309] In diese Richtung etwa auch Wischmeyer, AöR 2018, 1, 25.

8.1.4 Technische Standards

8.1.4.1 Protokollierung von Prozessen

Der Bericht skizziert den Stand der Forschung im Bereich Erklärbarkeit von ADM-Systemen. Diese Systeme sind eingebettet in komplexe Prozesse, in denen sie entwickelt und weiterentwickelt werden. Die Protokollierung der Abläufe ist notwendig, um das Gesamtverhalten des ADM-Systems zu verstehen.

Gerade für Algorithmen, die über sensible Lebensbereiche entscheiden, ist es durchaus denkbar, dass es standardisierte Prüfprotokolle, Anforderungslisten und Systembeschreibungen gibt, die während des Erstellungsprozesses angelegt werden müssen. In anderen Branchen ist es üblich, dass solche Dokumente vorhanden sind, insbesondere wenn es um den Schutz von Menschen und der Umwelt geht.

8.1.4.2 Entwicklung von Test- und Auditverfahren für ADM-Systeme

Das Testen von ADM-Systemen ist nach den Ergebnissen der Studie ein erfolgversprechendes Mittel zur Qualitätssicherung solcher Systeme und zum Schutz vor fehlerhaften algorithmischen Entscheidungen. Jedoch fehlt es derzeit weitgehend an anerkannten Test- und Auditverfahren.

Es ist daher dringend notwendig, die Entwicklung von Testverfahren für ADM-Systeme voranzutreiben, etwa durch entsprechende Forschungs- und Entwicklungsanstrengungen.

Weiterhin erforderlich ist die Festlegung qualitativer Standards für Testverfahren, da rechtliche Folgen nur an verlässliche Tests geknüpft werden können.

Damit eng verbunden ist auch die Auditierung der ADM-Systeme. Die Prüfprotokolle, die während einer Auditierung verwendet werden, sollten sich auf die jeweilige Domäne beziehen. Es müssen Erfahrungen gesammelt werden, damit dieses Auditing effizient durchgeführt werden kann.

Die Studie beschreibt ausführlich die Grundlagen und Herausforderungen des *Auditings* von ADM-Systemen und erörtert, wie diese Anwendung im Rahmen von Regulierung und Aufsicht finden können. Auditierung in Kombination mit Testverfahren, z.B. *Metamorphic Testing*, kann zur effizienten Überprüfung von ADM-Systemen eingesetzt werden.

8.1.4.3 Standardisierte Schnittstellen

ADM-Systeme brauchen wohldefinierte Schnittstellen, damit sie nach außen ohne großen Aufwand abgefragt werden können. Dies muss nicht notwendigerweise bedeuten, dass diese Schnittstellen für jeden offen und zugänglich sind. Im Falle eines Audits oder einer Überprüfung des Verhaltens durch einen Testdatensatz muss eine technische Schnittstelle zur Verfügung stehen. Um die oben erwähnte effiziente Abfrage zu ermöglich, ist es darüber hinaus notwendig, dass diese klar definiert ist.

Die Definition der Schnittstelle muss noch expliziter untersucht und ausgearbeitet werden. Einerseits muss sie konkret genug sein, damit sie hilfreich ist, andererseits sollte sie generisch sein, damit sie auch für zukünftige Anwendungen noch verwendet werden kann. Neben der technischen Standardisierung dieser Schnittstellen ist deren Bereitstellung und Mindestfunktionalität mit den Regulierungsmaßnahmen abzustimmen, zu verzahnen und

durch ein qualifiziertes Gremium anhand der wissenschaftlichen Entwicklung kontinuierlich zu verbessern und weiterzuentwickeln.

8.2 Organisatorische Maßnahmen

Für den Schutz gegen problematische algorithmische Entscheidungen sind tatsächliche Maßnahmen von großer Bedeutung, die hier in Abgrenzung zur Forschung und Gesetzgebung als „organisatorische" Maßnahmen bezeichnet werden. Wesentliche Maßnahmen sind die Aufklärung und Information (siehe Kapitel 8.2.1) und die Errichtung einer staatlichen Stelle mit Zuständigkeit für ADM-Systeme (siehe Kapitel 8.2.2).

8.2.1 Aufklärung, Information und Beratung

Mit fortschreitender technischer Entwicklung wird auch der Einsatz maschineller Lernverfahren immer einfacher; fertige Frameworks ermöglichen die erfolgreiche Anwendung solcher Verfahren auch mit allenfalls oberflächlichen Informatik- und Statistikkenntnissen. Die Vermeidung unrichtiger, weil beispielsweise diskriminierender Entscheidungen erfordert nach heutigem Stand jedoch Spezialwissen. Daher kommt der Vermittlung von Wissen eine entscheidende Rolle zu – sowohl betreffend die rechtlichen Anforderungen (soweit sie bereits ausreichend geklärt sind) als auch betreffend die technischen Umsetzungs- und Überprüfungsmöglichkeiten. Mit wachsendem Bewusstsein für die Problematik fehlerhafter algorithmischer Entscheidungen sehen wir auch eine wachsende Nachfrage nach entsprechenden Informationen, die es zu befriedigen gilt.

8.2.1.1 Aufklärung und Beratung als staatliche Aufgabe

Um dies zu erreichen, reicht es nicht aus, bei der Ausbildung anzusetzen; Informationen sollten auch kurzfristig zur Verfügung gestellt werden. Angesichts der wachsenden Bedeutung algorithmischer Entscheidungen sollte erwogen werden, dies zumindest teilweise als staatliche Aufgabe anzusehen. Ähnlich wie es die Datenschutzbeauftragten des Bundes und der Länder oder das Bundesamt für Sicherheit in der Informationstechnik in ihrem jeweiligen Aufgabenbereich tun, könnte auch die Beratung und Aufklärung über Probleme algorithmischer Entscheidungen durch staatliche Stellen übernommen werden. Auf die institutionellen Aspekte gehen wir in Kapitel 8.2.2 näher ein.

Der Einsatz von Algorithmen stößt derzeit noch auf viel Unwissen.[310] In den Medien werden viele Anwendungen beschrieben, aber das Bewusstsein und das Gewahrsein in der Bevölkerung sind noch kaum vorhanden. Hier ist wesentliche Informations- und Aufklärungsarbeit zu leisten. Des Weiteren wird es nötig sein, aufgrund fremder und eigener Forschungen Grundlagen für diese Zwecke zu erarbeiten bzw. aufzubereiten.

Wenn Regierungen, Behörden und Unternehmen ADM einsetzen, so stellt sich die Frage des gesellschaftlichen Entscheidungsprozesses. Grundsätzlich wird der Einsatz befürwortet, weil ADM konsistenter als Menschen arbeitet, mit der Komplexität besser umgehen kann und effizienter ist.

[310] Krüger/Lischka 2018.

8.2.1.2 Institutionelle Wahrnehmung von Aufklärung und Beratung

Die Erfüllung dieser Aufgabe bedarf einer institutionellen Absicherung. Ansonsten könnten unzureichende Ressourcen vorhanden sein, um diesen Prozess zu begleiten.

Die organisatorischen Aufgabenstellungen einer Institution wären folgende:

- Sammlung von Kompetenz
- Aufklärungsarbeit durch Website, Broschüren, FAQ und Fragemöglichkeiten
- Information der Öffentlichkeit durch Presseaussendungen, Veröffentlichungen, Workshops und Konferenzen
- Beratung bzw. Diskussion mit Gesetzgeber und Regierung
- Mitwirkung bei der Standardisierung
- Interaktion mit den Hochschulen, Ausbildung und professionelle Zertifizierung

Die institutionelle Umsetzung könnte als Forschungsinstitution oder als Behörde erfolgen. Als Beispiel kann das Bundesamt für Sicherheit in der Informationstechnik (BSI) angeführt werden. Diese ist als Behörde für IT-Sicherheit organisiert und hat die Kompetenz in IT-Sicherheit für den Bund sowie für Forschung, Standardisierung, Aufklärung und Information der Öffentlichkeit. Denkbar ist auch die Errichtung einer spezifischen staatlichen Stelle, etwa einer Agentur für ADM-Systeme (siehe Kapitel 8.2.8 und 8.3).

Andere Optionen wären die projektbezogene oder privatrechtliche Institutionalisierung, bei welcher bestehende Organisationen durch Projekte diese Aufgaben übernehmen (z.B. Allianz-Zentrum für Algorithmen). Vorteil wäre eine wesentlich flexiblere Struktur bei weitgehender öffentlicher Kontrolle durch die Definition von Projektzielen und Budgets. Diese Aufgabe könnten bestehende Ministerien übernehmen (z.B. das BMJV). Die Aufgaben der Projektnehmer wären weitgehend gleich; des Weiteren könnte eine intensivere Teilnahme am rechtspolitischen Diskurs erfolgen.

8.2.2 Staatliche Stelle für algorithmische Entscheidungen

Die Ergebnisse der Studie legen nahe, dass Bedarf an einer staatlichen Stelle besteht, die sich mit den spezifischen Herausforderungen algorithmischer Entscheidungen befasst. Auch in der gegenwärtigen Debatte wird die Einrichtung spezifischer staatlicher Stellen mit Bezug zu Algorithmen[311] oder „intelligenten Systemen“[312] gefordert.

Diese Schlussfolgerung wird auch in anderen Staaten gezogen. So werden in mehreren Ländern besondere Kommissionen oder Räte eingerichtet, die sich mit der Thematik befassen. Dabei sind aus rechtlicher Sicht stets die Aufgaben einer solchen Stelle, ihre Organisationsform sowie Befugnisse von Bedeutung. Nachfolgend sollen aus den Ergebnissen der Studie Schlussfolgerungen für die Aufgaben und Befugnisse der Stelle gezogen werden:

[311] Siehe dazu oben 7.1.1.5 und 7.1.1.7 mit entsprechenden Nachweisen.
[312] So etwa Wischmeyer, AöR 2018, 1, 63.

8.2.2.1 Aufgaben

Ein wesentlicher Befund der Studie liegt in der wiederkehrenden Erkenntnis, dass derzeit ein erheblicher Bedarf an Kenntnis über die Bedeutung und Risiken von ADM-Systemen allgemein wie algorithmischer Entscheidungen im Besonderen besteht. Ebenso fehlt es an gefestigter Kenntnis zu den Möglichkeiten, etwaigen Risiken durch staatliches Handeln, insbesondere organisatorische oder rechtliche Maßnahmen, zu begegnen.

Die damit vorrangige Aufgabe des Staates, den Kenntniserwerb zu fördern, besteht auch im Hinblick auf die Sicherung, Erweiterung und Bündelung der Kompetenz staatlicher Stellen in Bezug auf ADM und algorithmische Entscheidungen.

Da es bisher keine derartige Stelle gibt, liegt es nahe, eine spezifische staatliche Stelle zu errichten, die sich mit diesem Aufgabenfeld befasst.

Als Aufgaben der Agentur sollte insgesamt unter anderem Folgendes festgelegt werden:

- Koordination der Forschung zu Rechtsfragen des ADM
- Entwicklung und Standardisierung von Testverfahren für ADM
- Förderung der Zertifizierung von ADM-Verfahren
- Durchführung von Tests von ADM-Systemen
- Information und Beratung der Öffentlichkeit zu ADM
- Einrichtung einer Beschwerdestelle für ADM und algorithmische Entscheidungen
- Förderung und Betreiben von Streitbeilegungsmechanismen für algorithmische Entscheidungen

Diese Auflistung versteht sich nicht als abschließend. Weitere Aufgabenfelder können sich ergeben, bedürfen aber teilweise weiterer Prüfung. So bedarf es etwa sorgfältiger Überlegung, ob die Überprüfung konkreter algorithmischer Entscheidungen zum Aufgabenbereich der Stelle gehören soll. Insoweit sind in aller Regel auch Datenschutzaufsichtsbehörden zuständig, ebenso besteht zivilrechtlicher Rechtsschutz.

8.2.2.2 Eingriffsbefugnisse

Ob und inwieweit staatliche Stellen mit Eingriffsbefugnissen bei Algorithmen oder algorithmischen Entscheidungen auszustatten sind, ist derzeit nicht klar. In der aktuellen Diskussion wird teilweise die Etablierung einer staatlichen Aufsicht für Algorithmen gefordert. Dabei werden der Aufsicht unterschiedliche Aufgaben und Befugnisse zugeordnet. So wird die staatliche Aufsicht teilweise im Zusammenhang mit einer Vorabkontrolle genannt,[313] teilweise auch im Zusammenhang mit der Kontrolle von ADM-Systemen.[314] Soweit man dieses Instrument befürwortet, müssen auch entsprechende Eingriffsbefugnisse geschaffen werden. Da diese Studie die Möglichkeiten einer staatlichen Aufsicht für Algorithmen nicht umfassend untersucht, beschränken sich die Überlegungen zu Eingriffsbefugnissen auf die soeben genannten Aufgaben im Zusammenhang mit einer staatlichen Stelle.

Die staatliche Stelle muss entsprechend mit den für ihre Aufgaben notwendigen Eingriffsbefugnissen ausgestattet werden. Sie sollte daher, um Aufgaben zur Forschung, zur

[313] Siehe dazu oben Kapitel 7.1.1.5.
[314] Siehe dazu oben Kapitel 7.1.1.7.

Entwicklung von Testverfahren und zur Information und Beratung der Öffentlichkeit effektiv wahrnehmen zu können, die Befugnis erhalten, von Herstellern und Betreibern Auskünfte über Eigenschaften von ADM-Systemen und deren Einsatz zu verlangen.

Weiter sollte die Stelle mit der Befugnis zur Durchführung von Tests von ADM-Verfahren ausgestattet werden. Die Stelle sollte daher Tests anordnen und auch selbst durchführen können.

Ob die Stelle auch mit der Überprüfung einzelner algorithmischer Entscheidungen ausgestattet werden soll, sollte weiteren Untersuchungen vorbehalten sein, da insoweit auch zu prüfen ist, ob dies zum Aufgabenbereich der Stelle gehören soll.

8.2.2.3 Organisationsform

Die Organisationsform einer solchen staatlichen Stelle ist nicht Gegenstand der Studie. Es erscheint jedoch sinnvoll, als organisatorische Maßnahme eine staatliche Einheit, etwa in Form einer „Agentur für ADM-Systeme“, zu gründen, die Informations- und Beratungsfunktionen wahrnimmt, die Funktionsfähigkeit von Test- und Auditingverfahren sicherstellt und die einen Beitrag zur Versachlichung des öffentlichen Diskurses leistet.

Die Errichtung einer umfassend zuständigen staatlichen Stelle für Algorithmen/ADM-Systeme erscheint wenig realistisch, weil der Schwerpunkt der Fragen eher technischer Art ist. Jedoch ist es sinnvoll, für den Teilaspekt der rechtlichen Rahmenbedingungen von ADM-Verfahren und algorithmischer Entscheidungen eine Stelle zu schaffen, in der die Aufgaben gebündelt werden.

Denkbar erscheint etwa die Errichtung einer Bundesoberbehörde zur Steigerung der Transparenz im Bereich algorithmischer Entscheidungsverfahren. Ein Vorbild könnte die Einrichtung des Bundesamts für Sicherheit in der Informationstechnik (BSI) sein, das über viele Jahre vor allem Aufgaben im Bereich des Erkenntnisgewinns wahrnahm und erst in jüngster Zeit mit Eingriffsbefugnissen gegenüber Unternehmen und einem stark erweiterten Aufgabenbereich ausgestattet wurde. Das BSI nimmt heute Aufgaben im Bereich der Entwicklung von Maßnahmen, der Forschung, der Aufklärung und Information der Öffentlichkeit wahr, wirkt bei der Standardisierung und Zertifizierung mit und koordiniert die internationale Zusammenarbeit. Derartige Tätigkeiten werden auch bei ADM-Systemen und algorithmischen Entscheidungen anfallen.

8.3 Gesetzgebung

8.3.1 Gesetzgebungsbedarf

Ein Ergebnis der Studie lautet, dass grundsätzlich legislativer Handlungsbedarf bei ADM-Systemen und algorithmischen Entscheidungen besteht.

Jedoch sind der Umfang des Regelungsbedarfs und ebenso die Möglichkeiten der Gesetzgebung derzeit noch nicht deutlich absehbar. Dies beruht zum einen darauf, dass das Gefahrenpotenzial von ADM und algorithmischen Entscheidungen noch bei weitem nicht umfassend bekannt ist. Zum anderen bedürfen zahlreiche Rechtsfragen der Klärung. Dies betrifft teilweise rechtliche Grundlagenentscheidungen, wie die Zulässigkeit des *dynamic pricing* und etwaige Gleichbehandlungspflichten aufgrund strukturellen Ungleichgewichts. Es

betrifft aber auch etliche Auslegungsfragen zu bestehenden Instrumenten. So ist derzeit etwa die Bedeutung der DSGVO für algorithmische Entscheidungen unklar.

Die Handlungsempfehlungen für gesetzgeberische Akte beschränken sich daher auf solche Felder, in denen gegenwärtig Maßnahmen ergriffen werden können. Dies könnte auch in Form einer Selbst- oder Ko-Regulierung erfolgen.

8.3.2 Tests von ADM-Systemen

Die Durchführung von Tests von ADM-Systemen ist ein wesentliches Element des Schutzes gegen fehlerhafte algorithmische Entscheidungen. Daher sollten, wie im Einzelnen dargelegt, sowohl die Grundlagen von Tests und ihrer Durchführung als auch die Bedeutung von Testergebnissen rechtlich abgesichert werden.

8.3.2.1 Rechtliche Rahmenbedingungen für Testverfahren

Die Rahmenbedingungen für Tests von ADM-Systemen und für die Überprüfung algorithmischer Entscheidungen sollten auf spezifischer gesetzlicher Grundlage festgelegt werden. Zu den rechtlichen Anforderungen im Einzelnen besteht jedoch noch erheblicher Forschungsbedarf (siehe dazu oben Kapitel 7.2.2.2), so dass gesetzliche Maßnahmen erst nach Abschluss entsprechender Forschungsarbeiten ergriffen werden sollten.

8.3.2.2 Pflicht zur Durchführung von Tests

Ein wesentlicher Aspekt der rechtlichen Rahmenbedingungen sind Anreize zur Durchführung von Tests. Insoweit sollten folgende Maßnahmen getroffen werden:

8.3.2.2.1 Pflicht zur Durchführung von Tests vor dem Einsatz von ADM-Systemen

ADM-Systeme sollten vor ihrem Einsatz hinreichend auf Fehler, insbesondere Diskriminierung, geprüft werden. Um dies zu gewährleisten, sollte eine Pflicht zur Durchführung hinreichender Tests ausdrücklich gesetzlich vorgegeben werden, sobald der rechtliche Rahmen für geeignete Testverfahren gelegt ist.

8.3.2.2.2 Recht auf Durchführung von Tests

Ein Recht auf Durchführung von Tests sollte jedenfalls für die zu gründende staatliche Stelle („Agentur für ADM-Systeme“, dazu oben) geschaffen werden. Die Gewährung eines individuellen Anspruchs für Betroffene oder Verbraucherschutzorganisationen sollte geprüft werden; insoweit stehen den zu erwartenden Vorteilen auch erhebliche Risiken gegenüber.

8.3.3 Transparenz- und Informationspflichten

Transparenz und Information sind wichtige Schutzinstrumente gegen potentielle Gefahren durch algorithmische Entscheidungen. Daher sollte die Gewährung von Information auch durch rechtliche Mittel und entsprechende legislative Maßnahmen sichergestellt werden.

Im Einzelnen besteht noch erheblicher Klärungsbedarf. Folgende gesetzgeberische Maßnahmen können aber bereits jetzt empfohlen werden.

8.3.3.1 Meldepflichten

Soweit, wie hier angeraten, eine staatliche Stelle für ADM-Systeme eingeführt wird, sollten Betreiber von ADM-Verfahren, deren Betrieb Risiken für Persönlichkeitsrechte begründet, gesetzlich zur Meldung des Einsatzes an die Stelle verpflichtet werden. Hersteller sollten verpflichtet werden, das (erstmalige) Inverkehrbringen derartiger ADM-Systeme zu melden, einschließlich aussagekräftiger Angaben zu dem jeweiligen System.

8.3.3.2 Informationspflichten

Betreiber von ADM-Systemen, die von algorithmischen Entscheidungen betroffen sind, sollten verpflichtet sein über den Einsatz des Systems zu informieren, soweit ein Schutzbedarf besteht (siehe dazu oben Kapitel 7.3.2). Da ein solcher Schutz teilweise schon durch die DSGVO gewährleistet wird, ist zu klären, welche Lücken und Regelungsoptionen insoweit bestehen.

9 Literatur

9.1 Aufsätze, Monographien, Kommentare, Beiträge in Tagungsbänden

Adler, P., Falk, C., Friedler, S.A., Nix, T., Rybeck, G., Scheidegger, C., Smith, B., Venkatasubramanian, S. (2018): Auditing black-box models for indirect influence. *Knowledge and Information Systems*, *54*(1), 95-122.

Albarghouthi, A., D'Antoni, L., Drews, S., Nori, A. (2016): Fairness as a program property, *arXiv preprint arXiv:1610.06067.*

Andrews, R., Diederich, J., Tickle, A.B. (1995): Survey and critique of techniques for extracting rules from trained artificial neural networks. *Knowledge-based systems* 8, no. 6, 373-389.

Angwin, J., Larson, J., Mattu, S., Kirchner, L. (2016): Machine Bias, *ProPublica,* May 23, 2016 [https://bit.ly/1XMKh5R].

Auer-Reinsdorff, A., Conrad, I. (2016): Handbuch IT- und Datenschutzrecht, 2. Aufl. 2016.

Avery, R.B., Beeson, P.E., Calem, P.S. (1997): Using HMDA data as a regulatory screen for fair lending compliance. *Journal of Financial Services Research,* 11.1-2, 9-42.

Balboni, P., Dragan, T. (2018): Controversies and Challenges of Trustmarks: Lessons for Privacy and Data Protection Seals. In Privacy and Data Protection Seals. *TMC Asser Press*, The Hague.

Balkin, Jack M. (2017): Free Speech in the Algorithmic Society: Big Data, Private Governance, and New School Speech Regulation (September 9, 2017). UC Davis Law Review, (2018 Forthcoming); Yale Law School, *Public Law Research Paper No. 615*. Available at SSRN [https://bit.ly/2x2GJEy] oder [http://dx.doi.org/10.2139/ssrn.3038939].

Barr, E. T., Harman, M., McMinn, P., Shahbaz, Mu., Yoo, S. (2014): The Oracle Problem in Software Testing: A Survey. *IEEE Transactions on Software Engineering*. 41 (5), 507-525.

Bauchspies, R., Croissant, J., Restivo, S. (2005): Science, Technology, and Society: A Sociological Approach, *London: Wiley-Blackwell.*

Benjamin, S.M. (2012): Algorithms and Speech. *University of Pennsylvania Law Review 161, no. 6, 1445-1494.*

Berk, R., Hyatt, J. (2015): Machine Learning Forecasts of Risk to Inform Sentencing Decisions, *Federal Sentencing Reporter 27, No. 4, 222-228.*

Berk, R., Heidari, H., Jabbari, S., Kearns, M., Roth, A. (2017): Fairness in Criminal Justice Risk Assessments: The State of the Art. *arXiv preprint arXiv:1703.09207* [https://arxiv.org/pdf/1703.09207.pdf].

Berman, M. L. (2014): Manipulative marketing and the First Amendment. *Georgetown Law Journal, 103,* 497.

Bishop, C. M. (2006): Machine learning and pattern recognition. *Information Science and Statistics. Springer,* Heidelberg.

Borgesius, F. Z., Poort, J. (2017): Online Price Discrimination and EU Data Privacy Law. *Journal of Consumer Policy*, 40(3), 347-366.

Bracha, O., Pasquale, F. (2007): Federal Search Commission-Access, Fairness, and Accountability in the Law of Search. *Cornell Law Review*, *93*, 1149.

Bräutigam, P., Schmidt-Wudy, F. (2015): Das geplante Auskunfts- und Herausgaberecht des Betroffenen nach Art. 15 der EU-Datenschutzgrundverordnung, Ein Diskussionsbeitrag zum anstehenden Trilog der EU-Gesetzgebungsorgane, *Computer und Recht*, 56-63.

Breiman, L., Friedman, J. H., Olshen, R. A., Stone, C. J. (2017): Classification and regression trees, Routledge, Erstauflage 1984.

Brennan-Marquez, K. (2017): Plausible Cause: Explanatory Standards in the Age of Powerful Machines. *Vanderbilt Law Review*, *70*, 1249.

Calem, P.S., Longhofer, S.D. (2002): Anatomy of a fair lending exam: The uses and limitations of statistics. *The Journal of Real Estate Finance and Economics,* 24.3 (2002), 207-237 [https://doi.org/10.1023/A:1015264813969].

Calo, R. (2013): Consumer Subject Review Boards, *Stanford Law Review Online* 66, 97.

Calo, R. (2014): The case for a federal robotics commission, *Brookings Institution Center for Technology Innovation.*

Calo, R. (2017): Artificial Intelligence Policy: A Primer and Roadmap. *University of California, Davis Law Review, 51*, 399.

Cath, C., Wachter, S., Mittelstadt, B., Taddeo, M., Floridi, L. (2017): Artificial Intelligence and the 'Good Society': the US, EU, and UK approach. *Science and engineering ethics*, 1-24.

Cavoukian, A., Chibba, M. (2018): Privacy Seals in the USA, Europe, Japan, Canada, India and Australia. Privacy and Data Protection Seals. *TMC Asser Press*, The Hague.

Chander, A. (2016): The racist algorithm. *Michigan Law Review*, 115, 1023.

Citron, D. K., Pasquale, F. (2014): The scored society: due process for automated predictions. *Washington University Law Review*, 89, 1.

Citron, D. K. (2007): Technological due process. *Washington University Law Review*, 85, 1249.

Cook, R.E: (1997): The Lenders' "Other Regulator": Fair Lending Enforcement by the Department of Justice, *Federal Reserve Bank of San Francisco*.

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., Huq, A. (2017, August): Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 797-806. ACM.

Dal Pozzolo, A., Caelen, O., Le Borgne, Y., Waterschoot, S., Bontempi, G. (2014): Learned lessons in credit card fraud detection from a practitioner perspective, *Expert Systems with Applications* 41.10, 915-4928.

Dauner-Lieb, B., Langen, W. (Hrsg.) (2016): Bürgerliches Gesetzbuch, Bd. 2/2, Schuldrecht, 3. Aufl.

Desai, D. R., Kroll, J. A. (2017): Trust But Verify: A Guide to Algorithms and the Law. *Harvard Journal for Law and Technology*, 31 [https://bit.ly/2oSaww6].

Diakopoulos, N. (2014): Algorithmic-Accountability: the investigation of Black Boxes. *Tow Center for Digital Journalism* [https://bit.ly/2oVJ3tz].

Dormehl, L. (2014): The Formula: How Algorithms Solve All Our Problems… and Create More. New York: Perigee Books.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., Zemel, R. (2012): Fairness through awareness. *In Proceedings of the 3rd innovations in theoretical computer science conference, 214-226, ACM.*

Dwork, C., Mulligan, D. K. (2013): It's not privacy, and it's not fair. *Stanford Law Review Online*, 66, 35.

Dzida, B., Groh, N. (2018): Diskriminierung nach dem AGG beim Einsatz von Algorithmen im Bewerbungsverfahren. *Neue Juristische Wochenschrift* 2018, 1917-1922.

Edwards, L., Veale, M. (2017): Slave to the algorithm? Why a 'right to an explanation' is probably not the remedy you are looking for. *Duke Law and Technology Review*, 16(1), 1-65.

Ellis, E., Watson P. (2012): EU Anti-Discrimination Law, 2. Auflage, *Oxford European Union Law Library*.

Epping, V., Hillgruber, C. (Hrsg.) (2014): Beck'scher Online-Kommentar Grundgesetz. Beck: München (zitiert als BeckOK Grundgesetz).

Ernst, C. (2017): Algorithmische Entscheidungsfindung und personenbezogene Daten, *Juristenzeitung*, 1026-1036.

Ernst, H., Braunroth, A., Franke, B., Wascher, A. (2013): AGG, Nomos-BR, 2. Aufl., AGG § 19 Rn. 4.

Ezrachi, A., Stucke, M. E. (2017): Artificial intelligence & collusion: When computers inhibit competition. *University of Lillinois Law Review, 1775.*

Fayyad, U., Piatetsky-Shapiro, G., Smyth, P. (1996): From data mining to knowledge discovery in databases. AI magazine 17 (3): 37.

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., Venkatasubramanian, S. (2015, August): Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 259-268. ACM.

Finkelstein, M. O. (1980): The judicial reception of multiple regression studies in race and sex discrimination cases. *Columbia Law Review* 80, no. 4, 737-754.

Finlay, S. (2014): Predictive Analytics, Data Mining and Big Data. Myths, Misconceptions and Methods, Basingstoke: Palgrave Macmillan.

Finn, E. (2017): What Algorithms Want: Imagination in the Age of Computing. Cambridge, MA: MIT Press.

Friedman, B., Nissenbaum, H. (1996): Bias in Computer Systems, *ACM Transactions on Information Systems*, Vol. 14, No. 3, July 1996, 330-347.

Friedman, B., Brok, E., Roth, S.K., John Thomas (1996): Minimizing bias in computer systems. *ACM SIGCHI Bulletin 28*, no. 1, 48-51.

Füser, K. (2013): Intelligentes Scoring und Rating: moderne Verfahren zur Kreditwürdigkeitsprüfung. Springer-Verlag.

Gal, M. S., Elkin-Koren, N. (2017): Algorithmic Consumers. *Harvard Journal of Law & Technology*, *30*(2), 309.

Gellert, R., de Vries, K., de Hert, P. Gutwirth, S. (2013): A Comparative Analysis of Anti-Discrimination and Data Protection Legislations, in Custers, B., Calders, T., Schermer, B., Zarsky, T. (Hrsg.), *Discrimination and Privacy in the Information Society*, Springer, 61-89.

Gola, P. (Hrsg.) (2017): Datenschutz-Grundverordnung, VO (EU) 2016/679 Kommentar, *Beck, München* (zitiert als Gola, DSGVO).

Gola, P., Schomerus, R. (2015): BDSG: Bundesdatenschutzgesetz Kommentar, 12. Aufl.

Goodman, B., Flaxman, S. (2017): European Union Regulations on Algorithmic Decision Making and a Right to Explanation. *AI Magazine* 38.3, 50-58.

Goodman, B., Flaxman, S. (2016): EU regulations on algorithmic decision-making and a “right to explanation”. In *ICML workshop on human interpretability in machine learning* (WHI 2016), New York, NY [http://arxiv.org/abs/1606.08813v1].

Goodman, B. (2016): Discrimination, Data Sanitisation and Auditing in the European Union's General Data Protection Regulation. *European Data Protection Law Review*, 2.

Grabitz, E., Hilf, M., Nettesheim, M. (Hrsg.) (2017): Das Recht der Europäischen Union: EUV/AEUV, *Beck, München.*

Gunning, D. (2017): Explainable artificial intelligence (xai). *Defense Advanced Research Projects Agency (DARPA)*.

Hacker, P. (2018): Teaching Fairness to Artificial Intelligence: Existing and Novel Strategies against Algorithmic Discrimination under EU Law. *Common Market Law Review.*

Hardt, M., Price, E., Srebro, N. (2016). Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, 3.315-3.323.

Harrington, Jr., Joseph, E. (2017): Developing Competition Law for Collusion by Autonomous Price-Setting Agents [https://bit.ly/2x00Fc3].

Harkens, A. (2018): The ghost in the legal machine: algorithmic governmentality, economy, and the practice of law. *Journal of Information, Communication and Ethics in Society*, 16(1), 16-31.

Hastie, T., Tibshirani, R., and Friedman, J. (2009): *The elements of statistical learning: data mining, inference, and prediction.* 2nd ed. New York: Springer.

Heiden, I., Wersig, M. (2017): Preisdifferenzierung nach Geschlecht in Deutschland, Studie im Auftrag der Antidiskriminierungsstelle des Bundes [https://bit.ly/2NsSXAM].

Heidrich, J., Bauer, P., Krupka, D. (2018): Strukturen und Kollaborationsformen zur Vermittlung von Data-Literacy-Kompetenzen, *Arbeitspapier (Nr. 32), Hochschulforum Digitalisierung*, Mai 2018 [https://bit.ly/2O43XBB].

Helveston, M. N. (2015): Consumer Protection in the Age of Big Data. *Washington University Law Review*, 93, 859.

Hirsch, D. (2014): That's Unfair-Or Is It: Big Data, Discrimination and the FTC's Unfairness Authority. *Kentucky Law Review 103,* 345.

Hildebrandt, M. (2008): A vision of ambient law. *Regulating Technologies*, 75-191.

Introna, L., Nissenbaum, H. (1999): Sustaining the Public Good Vision of the Internet: The Politics of Search Engines. Center for the Arts and Cultural Policy Studies, Working Paper, 9, 1999.

Ishii, K. (2017): Comparative legal study on privacy and personal data protection for robots equipped with artificial intelligence: looking at functional and technological aspects. *AI & Society,* 1-25.

Jones, M. L. (2015): The ironies of automation law: tying policy knots with fair automation practices principles. *Vanderbilt Journal of Entertainment & Technology Law,* 18, 77.

Joseph, M., Kearns, M., Morgenstern, J. H., Roth, A. (2016): Fairness in learning: Classic and contextual bandits. In *Advances in Neural Information Processing Systems,* 325-333.

Joshi, A. (2018): India's Proposed Data Protection Framework-Recommendations on Individual Participation Rights, Whitepaper [https://bit.ly/2DRTLqP].

Kearns, M. (2018): Data Intimacy, Machine Learning, and Consumer Privacy. Penn Law CTIC Whitepaper, May 2018.

Kelley, R., Schaerer, E., Gomez, M., Nicolescu, M. (2010): Liability in robotics: an international perspective on robots as animals. *Advanced Robotics*, 24(13).

Kramer, A. D., Guillory, J. E., Hancock, J. T. (2014): Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences*, 111(24), 8788-8790.

Kroll, J. A., Barocas, S., Felten, E. W., Reidenberg, J. R., Robinson, D. G., Yu, H. (2016): Accountable algorithms. *University of Pennsylvania Law Review*, *165*, 633.

Krüger, J., Lischka K. (2018): Damit Maschinen den Menschen dienen. Lösungsansätze, um algorithmische Prozesse in den Dienst der Gesellschaft zu stellen – Arbeitspapier. *Bertelsmann-Stiftung*.

Kühling, J., Buchner, B., (Hrsg.) (2017): Datenschutz-Grundverordnung, Bundesdatenschutzgesetz: DSGVO/BDSG, *Beck, München.*

Kühling, J., Martini, M., Heberlein, J., Kühl, B., Nink, D., Weinzierl, Q., Wenzel, M. (2016): *Die Datenschutz-Grundverordnung und das nationale Recht*, Verlagshaus Monsenstein und Vannerdat, Münster.

Kühling, J. (2017): Neues Bundesdatenschutzgesetz – Anpassungsbedarf bei Unternehmen, *Neue Juristische Wochenschrift*, 1985-1990.

Kugelmann, D. (2016): Datenfinanzierte Internetangebote, *DuD - Datenschutz und Datensicherheit*, Springer, 566.

Kuner, C., Svantesson, D. J. B., Cate, F. H., Lynskey, O., Millard, C. (2017): Machine learning with personal data: is data protection law smart enough to meet the challenge? *International Data Privacy Law*, 7(1), 1-2.

Kurgan, L., Musilek, P. (2006): A survey of Knowledge Discovery and Data Mining process models. *The Knowledge Engineering Review.* Volume 21 Issue 1, March 2006, 1-24, Cambridge University Press, New York, NY, USA, doi: 10.1017/S0269888906000737.

Lachaud, E. (2017): The General Data Protection Regulation and the rise of certification as a regulatory instrument. *Computer Law & Security Review.*

Legrand, P. (1996): How to compare now. Legal Studies 16 (2): 232-242.

Le Sueur, A. (2016): Robot Government: Automated Decision-Making and its Implications for Parliament. In: Horen, A., Le Sueur, A. (Hrsg.): Parliament: Legislation and Accountability. Hart Studies in Constitutional Law.

Lepri, B., Oliver, N., Letouzé, E., Pentland, A., Vinck, P. (2017): Fair, Transparent, and Accountable Algorithmic Decision-making Processes. *Philosophy & Technology*, 1-17.

Lessmann, S., Baesens, B., Seow, H. V., Thomas, L. C. (2015): Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research*, 247(1), 124-136.

Levi-Faur, D. (2005): The Global Diffusion of Regulatory Capitalism. *The Annals of the American Academy of Political and Social Science*, 12-32.

Lowry, S., Macpherson, G. (1988): A Blot on the Profession. *British Medical Journal.* 296 (6623), 657.

Mancuhan, K., Clifton, C. (2014): Combating discrimination using bayesian networks. *Artificial intelligence and law* 22, no. 2, 211-238.

Mantelero, A. (2016): Personal data for decisional purposes in the age of analytics: From an individual to a collective dimension of data protection. *Computer Law & Security Review*, 32(2), 238-255.

Mariscal, G., Marban, O., Fernandez, C. (2018): A Survey of Data Mining and knowledge discovery process Models and methodologies. *The Knowledge Engineering Review*. Retrieved 30 April 2018.

Martini, M. (2017): Algorithmen als Herausforderung für die Rechtsordnung. *JuristenZeitung*, *72*(21), 1017-1025.

Martini, M., Nink, N. (2017): Wenn Maschinen entscheiden … – vollautomatisierte Verwaltungsverfahren und der Persönlichkeitsschutz, *Neue Zeitschrift für Verwaltungsrecht*, – Extra, 20/2017, 1-14.

Massaro, T. M., Norton H. (2015): Siri-ously? Free Speech Rights and Artificial Intelligence. *Norwestern University Law Review*, 110, 1.169-1.194.

Mehra, S. K. (2016): Antitrust and the Robo-Seller: Competition in the Time of Algorithms, *Minnesota Law Review*, 100 (2016), 1323-1375.

Miller, J. M. (2003). A typology of Legal transplants: using sociology, Legal History and Argentine examples to explain the transplant Process. The American Journal of Comparative Law, 51 (4), 839-886.

Mitchell, T. (1997): Machine Learning, *McGraw Hill*.

Mowshowitz, A., Kawaguchi, A. (2002). Bias on the Web. *Communications of the ACM*, 45(9), 56-60.

Noble, S. U. (2018): Algorithms of Oppression: How search engines reinforce racism. New *York University Press*.

O'Neil, C. (2016): Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy. *Crown Random House*, New York.

Odlyzko, A. (2009): Network neutrality, search neutrality, and the never-ending conflict between efficiency and fairness in markets. *Review of Network Economics*, 8(1), 40-60.

Ohlhausen, M. K., Okuliar, A.P. (2015): Competition, Consumer Protection, and the right [approach] to privacy. *Antitrust Law Journal* 80.1, 121.

Okuliar, A., Kamenir, E. (2017): Pricing Algorithms: Conscious Parallelism or Conscious Commitment? *Competition Policy International*.

Olzen, D., Looschelders, D., Schiemann, G. (2014): Staudingers Kommentar zum Bürgerlichen Gesetzbuch, Schuldrecht, §§ 241-243, Sellier - de Gruyter.

Paal, B., Pauly, D. (Hrsg.) (2018): Datenschutz-Grundverordnung, 2. Aufl., Beck: München.

Pagallo, U. (2010): The human master with a modern slave? Some remarks on robotics, ethics, and the law. *Ethicomp 2010: The "backwards, forwards and sideways" changes of ICT*, 397-404.

Pasquale, F. (2017): Toward a Fourth Law of Robotics: Preserving Attribution, Responsibility, and Explainability in an Algorithmic Society, *Ohio State Law Journal*, 78, 1243.

Pasquale, F. (2015): The black box society: The secret algorithms that control money and information. *Harvard University Press.*

Pasquale, F. (2010): Beyond innovation and competition: The need for qualified transparency in internet intermediaries. *Northwest University Law Review 104*, 105.

Pasquale, F., (2010): Reputation Regulation: Disclosure and the Challenge of Clandestinely Commensurating Computing, in *The Offensive Internet: Speech, Privacy, and Reputation*, ed. Levmore, S. and Nussbaum, M.C., Cambridge, MA: Harvard University Press.

Pedreshi, D., Ruggieri, S., Turini, F. (2008): Discrimination-aware data mining. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '08).* ACM, New York, NY, USA, 560-568 [https://doi.org/ 10.1145/1401890.1401959].

Quinlan, J. R. (1993): C4.5: Programs for Machine Learning. *Morgan Kaufmann Publishers.*

Raina, R., Madhavan, A., Ng, A.Y. (2009): Large-scale deep unsupervised learning using graphics processors. In *Proceedings of the 26th annual international conference on machine learning*, 873-880, ACM.

Rezac, M., Rezac, F. (2011): How to measure the quality of credit scoring models. *Finance a úver - Czech Journal of Economics and Finance*, 61 (5), 486.

Rice, L., Swesnik, D. (2013): Discriminatory effects of credit scoring on communities of color. *Suffolk University Law Rev*iew 46, 935.

Ridsdale, C., Rothwell, J., Smit, M., (2015): Strategies and Best Practices for Data Literacy Education: Knowledge Synthesis Report, Dalhousie University, Canada.

Rixecker, R., Säcker, F. J., Oetker, H. (Hrsg.) (2012): Münchener Kommentar zum BGB. (*Beck: München)* (zitiert als MüKoBGB).

Ribeiro, Ma. T., Singh, S., Guestrin, C. (2016): Why should I trust you? Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Rodrigues, R., Wright, D., Wadhwa, K. (2013): Developing a privacy seal scheme (that works). *International Data Privacy Law*, 3, 100-116.

Rohde, N. (2018): Gütekriterien für algorithmische Prozesse. Eine Stärken- und Schwächenanalyse ausgewählter Forderungskataloge. Arbeitspapier. Bertelsmann Stiftung.

Russell, S. J., Norvig, P. (2010): Artificial intelligence: a modern approach. Pearson Education Limited, Malaysia.

Sandvig, C., Hamilton, K., Karahalios, K., Langbort, C. (2014): Auditing algorithms: Research methods for detecting discrimination on internet platforms. *Data and discrimination: converting critical concerns into productive inquiry*, 1-23.

Schaerer, E., Kelley, R., Nicolescu, M. (2009): Robots as animals: A framework for liability and responsibility in human-robot interactions. *Robot and Human Interactive Communication*, September 2009. RO-MAN 2009. The 18th IEEE International Symposium.

Shneiderman, B. (2016): Opinion: The dangers of faulty, biased, or malicious algorithms requires independent oversight, *Proceedings of the National Academy of Sciences*, Nov 2016, 113 (48) 13538-13540.

Siegel, E. (2013): Predictive Analytics: The Power to Predict Who Will Click, Buy, Lie, or Die, Wiley.

Siems, M. M. (2005): Numerical comparative law: do we need statistical evidence in law in order to reduce complexity. *Cardozo Journal of International and Comparative Law*, 13, 521.

Simitis, S. (Hrsg.) (2014): Bundesdatenschutzgesetz, 8. Aufl. 2014.

Steiner, C. (2013): Automate This: How Algorithms Took Over Our Markets, Our Jobs, and the World. Portfolio.

Stucke, M.E., Ezrachi, A. (2017): How Digital Assistants Can Harm Our Economy, Privacy, And Democracy. *Berkeley Technology Law Journal*, 32(3).

Szollosy, M. (2017): Robots, AI, and the question of 'e-persons'-a panel at the 2017 Science in Public conference, 10-12 July 2017. *Journal of Science Communication,* 16(4).

Taeger, J. (2014): Anmerkung zu einer Entscheidung des BGH vom 28.01.2014 (VI ZR 156/13; MMR 2014, 489) - Zum Umfang einer von der Schufa zu erteilenden Auskunft, MMR 2014, S. 492-494.

Teubner, G. (1998): Legal irritants: good faith in British law or how unifying law ends up in new divergences. *The Modern Law Review,* 61.1, 11-32.

Tene, O., Polonetsky, J. (2012): Big data for all: Privacy and user control in the age of analytics. *Northwestern Journal of Technology and Intellectual Property*, 11, 239.

Thelisson, E., Padh, K., Celis, L. E (2017): Regulatory Mechanisms and Algorithms towards Trust in AI/ML. In IJCAI-17 Workshop on Explainable AI (XAI) (p. 53).

Tutt, A. (2017): An FDA for Algorithms, *Administrative Law Review* 69, 83.

Veale, M., Binns, R. (2017): Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data. *Big Data & Society*, *4*(2).

Veale, M., Van Kleek, M., Binns, R. (2018, April): Fairness and Accountability Design Needs for Algorithmic Support in High-Stakes Public Sector Decision-Making. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*. ACM.

Wachter-Boettcher, S. (2017): Technically Wrong: Sexist Apps, Biased Algorithms, and Other Threats of Toxic Tech. *WW Norton & Company.*

Waltl, B., Vogl, R. (2018): Explainable AI or how to prepare Legal Informatics for the Next Wave of Artificial Intelligence, IRIS: Internationales Rechtsinformatik Symposium, Salzburg, Austria.

Wang, Z., Wang, X. (2017): Big Data in healthcare in China: Applications, obstacles, and suggestions. Frontiers in Data Science, 355.

Watson, A. (1996): Aspects of reception of law. *The American Journal of Comparative Law* 44 (2): 335-351.

Webber, M., Car, T. (2016): Looking at Europe from the USA: Current perspectives on data protection. *Journal of Data Protection & Privacy* 1 (1), 76-88.

Williams, R., Edge, D. (1996): The Social Shaping of Technology, *Research Policy* 25, 856-899.

Wischmeyer, Thomas (2018): Regulierung intelligenter Systeme, *Archiv des öffentlichen Rechts*, 143, 1-66.

Wolff, H. A., Brink, S (Hrsg.) (2013): Beck'scher Online-Kommentar Datenschutzrecht. (*Beck, München)* (zitiert als BeckOK DatenschutzR).

Wolfie, C. (2017): How Companies Use Personal Data Against People: Automated Disadvantage, Personalized Persuasion, and the Societal Ramifications of the Commercial Use of Personal Information; *Cracked Labs Working paper*, Vienna.

Zhao, J., Wang, T., Yatskar, M., Ordonez, V., Chang, K. W. (2017): Men also like shopping: Reducing gender bias amplification using corpus-level constraints. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP) 2017*, 2941-2951, ACL 2017 [https://bit.ly/2QhofsX]. arXiv preprint arXiv:1707.09457.

Žliobaitė, I., Custers B. (2016): Using sensitive personal data may be necessary for avoiding discrimination in data-driven decision models. *Artificial Intelligence and Law* 24, no. 2, 183-201.

9.2 Regierungs- und Konsultationsdokumente, Berichte, Urteile

ACLU Press Release: ACLU Challenges Computer Crimes Law That is Thwarting Research on Discrimination, Online Aug. 2016.

American Bankers Association: More Really Is Less: The Data Fog Frustrates HMDA, April 2017, https://www.aba.com/Advocacy/Documents/HMDAWhitePaper2017.pdf.

Anhörungen der United States House of Representatives Committee on Energy and Commerce Subcommittee on Digital Commerce and Consumer Protection: „Algorithms: How Companies' Decisions About Data and Content Impact Consumers", November 29, 2017.

Australian Government Productivity Commission: Productivity Commission Draft Report: Data Availability and Use, October 2016 [https://bit.ly/2QkzJM2].

BAG, Urteil vom 27. Januar 2011 - 8 AZR 483/09 -, juris.bundesarbeitsgericht.de.

BT-Drucksache 16/1780: Entwurf eines Gesetzes zur Umsetzung europäischer Richtlinien zur Verwirklichung des Grundsatzes der Gleichbehandlung, 8. Juni 2006 [https://bit.ly/2yMWL4e].

Board Of Governors Of The Federal Reserve System Washington, Sr 11-7: Guidance on Model Risk Management, 4.4.2011.

Bryson, J. Testimony for the House of Lords Select Committee on Artificial Intelligence.

Bundesanstalt für Finanzdienstleistungsaufsicht (2018): Big Data trifft auf künstliche Intelligenz. Herausforderungen und Implikationen für Aufsicht und Regulierung von Finanzdienstleistungen.

BVerfG, Beschluss des Zweiten Senats vom 18. Juni 2008 - 2 BvL 6/07 -, BVerfGE 121, 241.

BVerfG, Beschluss des Ersten Senats vom 11. April 2018 - 1 BvR 3080/09 -, www.bverfg.de.

Chopin and Germaine, „A comparative analysis of non-discrimination law in Europe 2015" (Report for DG Justice and Consumers, 2016.

EuGH, Urteil vom 6. Dezember 2007, Rs. C-300/06, Slg. 2007, I-10573 – Voß.

Executive Office of the President of the United States (Council of Economic Advisors): Big Data and Differential Pricing, 2015 [https://bit.ly/2xRYC7R].

FAT/ML Principles for Accountable Algorithms and a Social Impact Statement for Algorithms.

Federal Trade Commission Forum, fintech Series: Marketplace Lending, June 9, 2016.

Federal Trade Commission, Protecting Consumer Privacy in an Era of Rapid Change: Recommendations for Businesses and Policymakers (Mar. 2012).

Financial Conduct Authority: Feedback Statement FS16/5: Call for Inputs on Big Data in retail general insurance, September 2016 [https://bit.ly/2NtzT5o].

Gosselin, S., Jones, A. and Martin A. (2017): Are Antitrust Laws Up to the Task? A US/EU Perspective on Anti-Competitive Algorithm Behavior, Hausfeld LLP [https://bit.ly/2CCu8hu].

Grutter v. Bollinger, 539 U.S. 306 (2003).

McSweeny, T., Commissioner of the Federal Trade Commission: competition law: keeping pace in a digital age 15.4.2016.

Milieu Ltd., „Comparative study on access to justice in gender equality and anti-discrimination law" (Report for DG Justice, 2011) [https://bit.ly/2x0dhiH].

New York City Council: A Local Law to amend the administrative code of the city of New York, in relation to automated processing of data for the purposes of targeting services, penalties, or policing to persons.

OECD Directorate For Financial And Enterprise Affairs Competition Committee: Algorithmic Collusion: Problems and Counter-Measures.

Office of Fair Trading (2010). Online targeting of advertising and prices. A market study [https://bit.ly/2wX4EWY].

Pasquale, F., written Testimony of Before the United States Senate Committee on Banking, Housing, and Urban Affairs, „Exploring the Fintech Landscape".

Pasquale, F., Before the United States House of Representatives Committee on Energy and Commerce Subcommittee on Digital Commerce and Consumer Protection [https://bit.ly/2MiBgPI].

Retail Credit Co. of Atlanta, Ga: Hearing Before a Subcommittee of the Committee on Government Operations, 90th Cong. 44-45 (May 16, 1968).

Sullivan, D., A Letter to the FTC Regarding Search Engine Disclosure Compliance.

UK All-Party Parliamentary Group on Artificial Intelligence (APPG AI): Evidence meeting, 23.1.2018.

UK Government, All-Party Parliamentary Group on Artificial Intelligence [APPG AI]: Ethics and Legal in AI: Decision Making and Moral Issues, 27.3.2017.

UK Government, Department for Department for Digital, Culture, Media & Sport and Department for Business, Energy & Industrial Strategy: Growing the artificial intelligence industry in the UK Oct 2017.

UK House of Commons, Science and Technology Committee (Commons)Algorithms in decision-making inquiry, 23.1.2018.

UK House of Commons Science and Technology Committee: Commission for Robotics and artificial intelligence, Oct 2016.

UK Information Commissioner's Office: Royal Free – Google Deepmind trial failed to comply with data protection law, July 3 (2017).

UK Information Commissioner's Office: The Information Commissioner's Office's (ICO's) response to the Science and Technology Committee's call for evidence on algorithms in decision-making, 2017 [https://bit.ly/2CEpKhV].

UK Information Commissioner´s Office: Big data, artificial intelligence, machine learning and data protection, Version 2.2, 2017 [https://bit.ly/2mF1kLj].

US Executive Office of the President and Jason Furman, John P. Holdren, Cecilia Muñoz, Megan Smith and Jeffery Zients, „Artificial Intelligence, Automation, and the Economy", Technical report, National Science and Technology Council, Washington D.C. 20502, October 2016.

US Department of Justice, „Algorithms and Collusion - Note by the United States", OECD Background Paper, Roundtable on Algorithms and Collusion, DAF/COMP/WD(2017)41, June 2017.

US Federal Trade Commission, Big Data: A Tool For Inclusion or Exclusion (Jan. 2016).

US Public Policy Council (USACM) Statement on Algorithmic Transparency and Accountability.

US Senate: 115th Congress 1st Session: To require the Secretary of Commerce to establish the Federal Advisory Committee on the Development and Implementation of Artificial Intelligence, and for other purposes.

New Zealand Human Rights Commission: Privacy, Data and Technology: Human Rights Challenges in the Digital Age, May 2018 [https://bit.ly/2CB4iKY].

Ohlhausen, Maureen: „Should We Fear the Things That Go Beep in the Night? Some Initial Thoughts on the Intersection of Antitrust Law and Algorithmic Pricing“, Remarks by the Acting Chairman of the U.S. Federal Trade Commission at the “Concurrences Antitrust in the Financial Sector Conference”, New York, May 23, 2017.

Vestager, M. (2017): Algorithms and Competition, Remarks by the European Commissioner for Competition at the Bundeskartellamt 18th Conference on Competition, Berlin, March 16, 2017.

World Wide Web Foundation: Algorithmic Accountability: Applying the concept to different country contexts, July 2017 [https://bit.ly/2O50iDC].

Autoren

Folgende Experten sind an der Erstellung dieser Studie beteiligt (in alphabetischer Reihenfolge der Nachnamen):

Prof. Dr. Georg Borges ist Inhaber des Lehrstuhls für Bürgerliches Recht, Rechtsinformatik, deutsches und internationales Wirtschaftsrecht sowie Rechtstheorie der Universität des Saarlandes. Er ist Gründungsmitglied und Sprecher des Vorstands der Arbeitsgruppe Identitätsschutz im Internet e.V. (a-i3), Vorstandsmitglied des Horst Görtz Instituts für Sicherheit in der Informationstechnik der Ruhr-Universität Bochum und ehemaliger Richter am Oberlandesgericht Hamm.

Dr. Matthias Grabmair ist Systems Scientist am Language Technologies Institute der Carnegie Mellon University in Pittsburgh, USA. Er lehrt in einem Masterprogramm in Computational Data Science und forscht im Bereich der Anwendung von Techniken aus Artificial Intelligence, Natural Language Processing und Machine Learning auf juristische Texte und Daten, insbesondere an berechenbaren Modellen juristischer Argumentation. Er ist Mitglied im Editorial Board des Journal of Artificial Intelligence and Law.

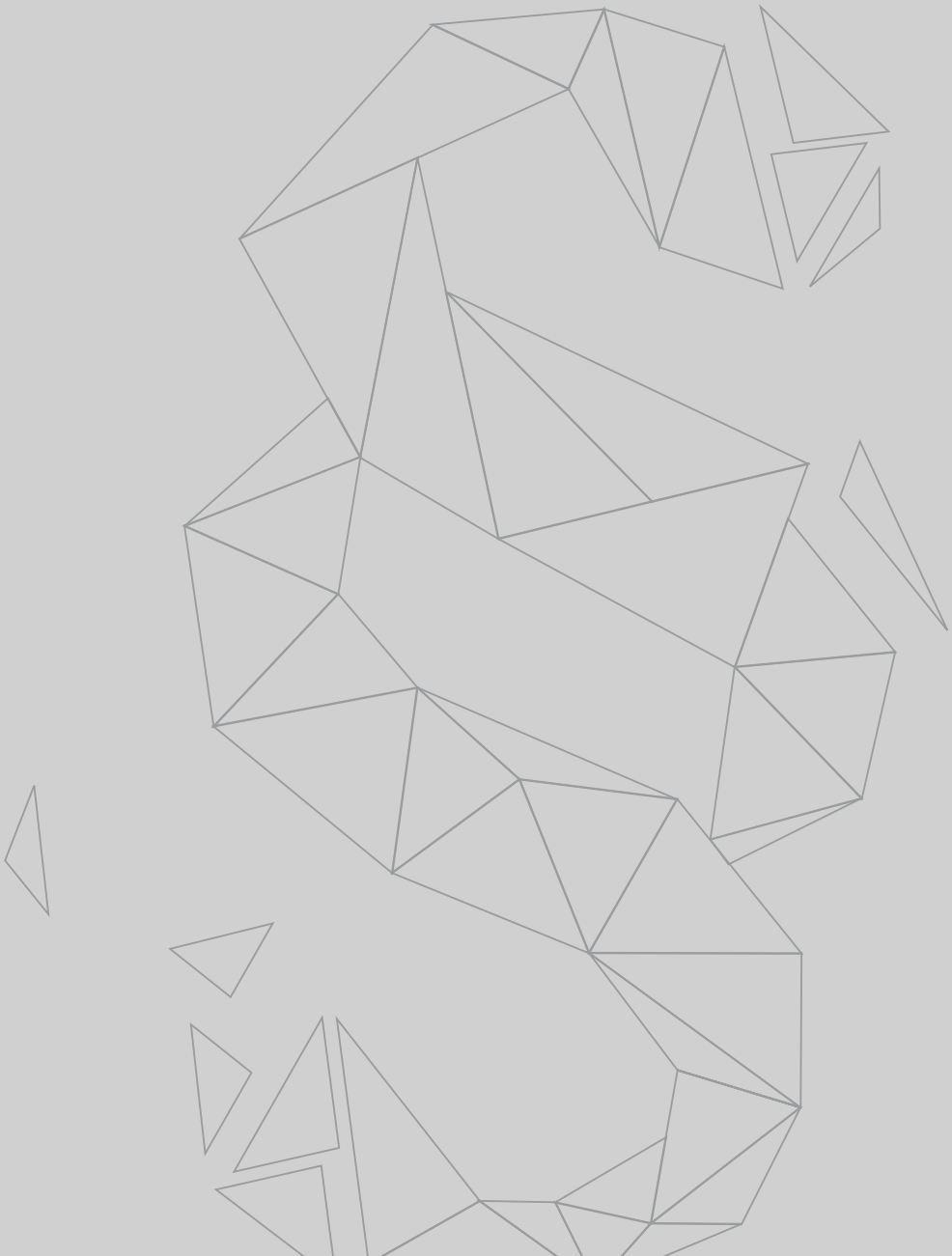
Daniel Krupka ist Geschäftsführer der Gesellschaft für Informatik und Leiter der Berliner Geschäftsstelle. Er verantwortet die Kommunikation, Öffentlichkeitsarbeit sowie den Kontakt zu Politik, den Bundesministerien und den Partnern auf nationaler, europäischer und internationaler Ebene.

Prof. Burkhard Schäfer ist Inhaber des Lehrstuhls Computational Legal Theory an der Edinburgh Law School, University of Edinburgh. Er ist Mitglied der Management group of the RCUK funded CREATE research network on the future of copyright in the digital economy und hat mehr als 90 wissenschaftliche Beiträge in den Bereichen „legal expert system design“, Semantic Web und Regulierung neuer Technologien veröffentlicht.

Prof. Mag. Dr. Dr. Erich Schweighofer ist außerordentlicher Universitätsprofessor an der Universität Wien, Lehr- und Forschungstätigkeit in den Fächern Rechtsinformatik, Völkerrecht und Europarecht, Leiter der Arbeitsgruppe Rechtsinformatik, Juridicum, Universität Wien und Hauptorganisator des Internationalen Rechtsinformatik Symposions IRIS [www.univie.ac.at/RI/IRIS2018]. Er ist Sprecher der Fachgruppe Rechtsinformatik der GI bzw. OCG und Mitglied bei CEPIS LIS. Neben seiner Universitätstätigkeit hat er langjährige Erfahrung in der Verwaltung (Außenministerium, Verwaltungsakademie des Bundes, Europäische Kommission).

Prof. Dr.-Ing. Christoph Sorge ist Professor an der Rechtswissenschaftlichen Fakultät und kooptierter Professor im Fachbereich Informatik der Universität des Saarlandes. Er ist Senior Fellow des Deutschen Forschungsinstituts für Öffentliche Verwaltung Speyer, Autor von über 60 wissenschaftlichen Veröffentlichungen in Informatik und Rechtswissenschaft und Vorstandsmitglied des Deutschen EDV-Gerichtstags.

Bernhard Waltl ist wissenschaftlicher Mitarbeiter der TU München und forscht seit einigen Jahren an der Schnittstelle zwischen Recht und Informatik mit Schwerpunkt Künstlicher Intelligenz und Datenanalyse. Er ist Mitgründer des interdisziplinären Forschungsprogramms „Lexalyze“ und arbeitet intensiv mit Juristen und Rechtswissenschaftlern zusammen.



Sachverständigenrat für Verbraucherfragen

Der Sachverständigenrat für Verbraucherfragen ist ein Beratungsgremium des Bundesministeriums der Justiz und für Verbraucherschutz (BMJV). Er wurde im November 2014 vom Bundesminister der Justiz und für Verbraucherschutz, Heiko Maas, eingerichtet. Der Sachverständigenrat für Verbraucherfragen soll auf der Basis wissenschaftlicher Erkenntnisse und unter Berücksichtigung der Erfahrungen aus der Praxis das Bundesministerium der Justiz und für Verbraucherschutz bei der Gestaltung der Verbraucherpolitik unterstützen.

Der Sachverständigenrat ist unabhängig und hat seinen Sitz in Berlin.

Vorsitzende des Sachverständigenrats ist Prof. Dr. Lucia Reisch.