Zeitschrift für Praktische Philosophie
Band 7, Heft 1, 2020, S. 265–300
www.praktische-philosophie.org
https://doi.org/10.22613/zfpp/7.1.9

Intransparente Diskriminierung durch maschinelles Lernen

Intransparent discrimination by machine learning

Heiner Koch, Duisburg-Essen

Zusammenfassung: Maschinelles Lernen kann zu Diskriminierungen anhand von völlig neuen Merkmalen führen. Diese Merkmale können uns außerdem nicht transparent sein. Um dies zu zeigen, werde ich argumentieren, dass Diskriminierung nicht nur anhand von Merkmalen in abschließenden Listen erfolgen kann, sondern grundsätzlich anhand beliebiger Merkmale, und dass Algorithmen trotz fehlender mentaler Eigenschaften diskriminierend sein können. Zudem werde ich die Probleme beschreiben, die damit einhergehen, dass diese neuen Formen der Diskriminierung nur schwer erkennbar sind, weil die eingesetzten Algorithmen intransparent sind. Hierbei sind drei Arten der Intransparenz zu unterschieden, die jeweils unterschiedliche Auswirkungen auf die Erkennbarkeit von Diskriminierungen haben:

(i) Merkmale, anhand derer ungleich behandelt wird, können unbekannt sein
(ii) Merkmale, anhand derer ungleich behandelt wird, können unverständlich sein (sie sind zu komplex oder „chaotisch", um von Menschen sinnvoll erfasst werden zu können)
(iii) es können Erklärungen dafür fehlen, wie und weshalb bestimmte Merkmale für eine Ungleichbehandlung herangezogen werden

Gerade die Kombination aus Intransparenz und Diskriminierung anhand neuer Merkmale stellt eine besondere Herausforderung für die philosophische Debatte um Diskriminierung dar. Bisherige Ansätze beschränken sich darauf, verständliche Merkmale zu identifizieren oder zu erzeugen, die ihre Träger_innen unter expliziten Diskriminierungsschutz stellen. Damit geraten jedoch neue Diskriminierungen aus dem Blick, die in Zukunft erhebliche Auswirkungen haben können. Um intransparente Diskriminierungen aufzudecken, müssen unbekannte Merkmale identifiziert werden, unverständliche Merkmale hinreichend verständlich gemacht werden und Korrelationen und Trainingsverfahren erklärbar sein. Anschließend müssen die diskriminierenden Elemente des Algorithmus beseitigt werden können. Insofern keine angemessenen Lösungsstrategien vorliegen, muss darüber nachgedacht werden, für

bestimmte Kontexte nur Formen des maschinellen Lernens zuzulassen, die sich hinreichend auf diskriminierende Konsequenzen untersuchen lassen.

Schlagwörter: Diskriminierung, maschinelles Lernen, Algorithmen, Explainable Artificial Intelligence, statistische Diskriminierung, Intransparenz

Abstract: Machine Learning can lead to discrimination based on new features that are intransparent. In order to show this, I will argue that discrimination cannot only occur based on features in exhaustive lists but on any feature and that algorithms can be discriminating in spite of a lack of mental attributes. I will show some problems that come with new forms of discrimination that are not easily detectable because of the intransparency of the algorithms. Three forms of intransparency that have different effects on the possibility to detect discriminations have to be distinguished:
(i) features of differential treatment can be unknown
(ii) features of differential treatment can be unimaginable (they are too complex or "chaotic" to be grasped by humans)
(iii) explanations for the use of certain features for differential treatment could lack
Especially the combination of intransparency and new features of discrimination is challenging for the philosophical debate about discrimination. Previous solutions were restricted to identify or create imaginable features that are explicitly protected by anti-discrimination law. New forms of discrimination that could have considerable impact in the future are neglected. In order to identify intransparent discriminations unknown features have to be identified, unimaginable features have to be made sufficiently imaginable and correlations and training methods have to be explainable. Then it has to be possible to remove discriminating elements. Insofar no suitable solutions can be found, it must be considered whether we want to allow for specific contexts only those machine learning methods which can sufficiently be examined for discriminating consequences.

Keywords: discrimination, machine learning, algorithms, explainable artificial intelligence, statistical discrimination, intransparency

1. Einleitung

Algorithmen entscheiden in einem zunehmenden Maße über unser Leben. Sie treffen eine Vorauswahl bei Jobbewerbungen, schätzen unsere Kreditwürdigkeit ein, halten uns für ein Sicherheitsrisiko, bestimmen Versicherungstarife oder bewerten unsere Arbeitsleistung. Subtiler ist ihr Einfluss etwa bei automatisierten Kaufempfehlungen, Werbeeinblendungen, der Erzeugung von Feeds in sozialen Netzwerken oder dem Vorschlag von Aktivitäten. Um dies alles tun zu können, benötigt der Algorithmus personen-

bezogene Daten, aus denen Profile und statistische Einschätzungen erzeugt werden. Damit sind die Unterscheidungen, die der Algorithmus vornimmt, anfällig für problematische Diskriminierungen.[1] Dies ist bekannt und mittlerweile auch intensiv untersucht worden (Custers, Calders, Schermer und Zarsky 2013). So können etwa Schwarze durch maschinelles Lernen als weniger kreditwürdig eingeschätzt (Fuster, Goldsmith-Pinkham, Ramadorai und Walther 2018) oder von Gesichtserkennungssoftware nicht oder falsch erkannt werden (Buolamwini und Gebru 2018). Frauen können bei der automatisierten Vorsortierung von Bewerbungen als weniger qualifiziert bewertet werden (Dastin 2018). Muslime oder Menschen aus mehrheitlich muslimischen Ländern können als Sicherheitsrisiko eingestuft werden (Mozur 2019). Hierbei handelt es sich oft, wenn auch nicht immer, um nicht intendierte Diskriminierungen. Der Algorithmus lernt anhand eines mit historischen Diskriminierungen vorbelasteten Datensatzes, wie Klassifizierungen oder Bewertungen vorzunehmen sind. Die diskriminierenden Elemente in dem Datensatz werden anschließend reproduziert. Wenn ein Unternehmen bisher bevorzugt Männer eingestellt hat, kann es sein, dass der Algorithmus lernt, dass Männer für die Jobs besser geeignet sind. Nicht immer hängt der diskriminierende Effekt jedoch daran, dass der Trainingsdatensatz in dieser Art vorbelastet ist. Der Algorithmus kann auch anhand von Merkmalen lernen zu diskriminieren, anhand derer bisher keine Diskriminierung oder auch nur eine Differenzierung stattgefunden hat – etwa anhand von Kontakten in sozialen Netzwerken, Stimmanalysen oder Bewegungsprofilen.

Bei der Suche nach systematischen Ungleichbehandlungen – und damit potentiellen Diskriminierungen – wird für gewöhnlich nach Benachteiligungen von denjenigen Gruppen gesucht, deren Diskriminierung verboten oder zumindest gesellschaftlich stark geächtet ist (z. B. FRA 2018). Hier soll jedoch argumentiert werden, dass damit ein zunehmend wichtigerer Bereich vernachlässigt wird: die Benachteiligung von Gruppen, die nicht explizit verboten ist. Ein Algorithmus könnte etwa erlernen, bei der Kreditvergabe systematisch Menschen zu benachteiligen, die einen niedrigen Bildungsabschluss besitzen oder in sozialen Netzwerken mit Menschen befreundet

1 Damit ist nicht gesagt, dass menschliche Entscheidungen weniger diskriminierungsanfällig sind. Oft kann ein Algorithmus auch eingesetzt werden, um fehleranfällige und diskriminierende menschliche Entscheidungen zu ersetzen. Dies funktioniert jedoch dann besonderes gut, wenn der Algorithmus so transparent und beeinflussbar ist, dass die entsprechenden Diskriminierungen verhindert werden können.

sind, die über ein schlechtes Kreditscoring verfügen. Wann solche nicht explizit verbotenen Formen der Benachteiligung eine Diskriminierung darstellen, ist umstritten. Hierbei handelt es sich zwar um keine Besonderheit, die erst im Zusammenhang mit maschinellem Lernen auftritt. Dennoch gibt es hier aufgrund der Intransparenz oder der fehlenden Interpretierbarkeit des maschinellen Lernens einige neue Aspekte, die beachtet werden sollten, um Gefahren der Diskriminierung im Zusammenhang mit maschinellem Lernen angemessen begegnen zu können. Da nicht unbedingt bekannt ist, was genau der Algorithmus gelernt hat, ist es auch nicht unbedingt bekannt, anhand welcher Merkmale Klassifikationen und Bewertungen vorgenommen werden. Der Algorithmus erscheint als eine Black Box, die anhand von unbekannten Merkmalen und Verfahren (dies schließt den Lernprozess des Algorithmus mit ein) von einem Input zu einem Output kommt. Daher muss oft der Output des Algorithmus erst unabhängig von explizitem Wissen über die Black Box untersucht werden, bevor festgestellt werden kann, ob der Algorithmus systematisch bestimmte Gruppen benachteiligt. Auch die Untersuchung der Trainingsdaten kann aufschlussreich sein.

Nun sind drei Arten der Intransparenz zu unterschieden, die jeweils unterschiedliche Auswirkungen auf die Erkennbarkeit von Diskriminierungen haben:

(i) Merkmale, anhand derer ungleich behandelt wird, können unbekannt sein
(ii) Merkmale, anhand derer ungleich behandelt wird, können unverständlich sein (sie sind zu komplex oder „chaotisch“, um von Menschen sinnvoll erfasst werden zu können)
(iii) es können Erklärungen dafür fehlen, wie und weshalb bestimmte Merkmale für eine Ungleichbehandlung herangezogen werden

Aufgrund dieser Intransparenzen ist es nicht mehr so einfach möglich, gezielt im Output, also dem Ergebnis des maschinellen Berechnungsverfahrens, nach möglicherweise problematischen Merkmalen der Ungleichbehandlung zu suchen. Stattdessen muss sehr allgemein nach Merkmalen der Ungleichbehandlung gesucht werden. Diese Merkmale müssen außerdem mitunter erst verständlich gemacht werden, um mit diesen umgehen zu können. Und um beurteilen zu können, ob eine Ungleichbehandlung sachlich und normativ angemessen oder diskriminierend ist, müssen (zumindest manchmal) verständliche Erklärungen für die Ungleichbehandlung durch den Algorithmus gefunden werden. Forschung im Bereich der Explainable Artificial Intelligence (XAI) wird dabei aktuell intensiv betrieben.

Ist es beispielsweise bekannt, dass eine Benachteiligung bei der Kreditvergabe aufgrund eines niedrigen Bildungsabschlusses stattfindet, so lassen sich Erwägungen darüber anstellen, ob diese Benachteiligung sachgerecht und normativ angemessen ist oder nicht. Diese Art der Erwägung ist bei intransparenten Benachteiligungen durch maschinelles Lernen nicht möglich. Unbekanntes oder Unverständliches kann kaum in Erwägungsprozesse einbezogen werden. So identifizieren Custers et al. (2013) die Notwendigkeit diesen Bereich in Zukunft intensiver zu erforschen. Sie weisen ebenso auf einige besondere Probleme hin, die mit dieser neuen Form der Diskriminierung von neuen Gruppen verbunden sind: „These new groups might be dispersed throughout society. Thus, they will lack the minimal political force to bring the issues of their misfortune to the forefront of the legal discussion. Even worse, given the inherent obscurity of the data mining practices [...] those adversely impacted by these processes might not even know this is happening!“ (353). Dennoch findet sich etwa in dem umfassenden Sammelband „The Routledge Handbook of the Ethics of Discrimination“, 2018 von Lippert-Rasmussen herausgegeben, kein Artikel zur Diskriminierung durch Algorithmen, maschinelles Lernen oder den neuen Formen der Diskriminierung, die durch maschinelles Lernen entstehen können.

Um für die These zu argumentieren, dass maschinelles Lernen zu Diskriminierungen anhand völlig neuer und intransparenter Merkmale führen kann, werde ich im Anschluss an allgemeine Ausführungen zum Diskriminierungsbegriff (2.) zeigen, dass Diskriminierung nicht nur anhand von Merkmalen in abschließenden Listen erfolgen kann, sondern grundsätzlich anhand beliebiger Merkmale (3.1), und dass Algorithmen trotz fehlender mentaler Eigenschaften diskriminierend sein können (3.2). In Teil 4 werde ich die Probleme beschreiben, die damit einhergehen, dass diese neuen Formen der Diskriminierung nur schwer erkennbar sind, weil die eingesetzten Algorithmen intransparent sind.

2. Diskriminierungsbegriff

Im Folgenden soll es darum gehen, einige Entscheidungen bezüglich der weiteren Verwendung des Diskriminierungsbegriffs explizit zu machen und begriffliche Klärungen vorzunehmen. Erst in Teil 3 soll es darum gehen, substantieller für einen bestimmten Diskriminierungsbegriff zu argumentieren, soweit dies nötig ist, um das Phänomen der Ungleichbehandlung durch maschinelles Lernen anhand neuer Merkmale als Diskriminierung fassen zu können.

In verschiedenen Ansätzen (etwa Lippert-Rasmussen 2014 oder Eidelson 2015) wird zunächst ein moralisch neutraler Diskriminierungsbegriff entwickelt (anders etwa Wasserman 1998), auf dem anschließend ein gehaltvollerer, von normativen Überlegungen angeleiteter Diskriminierungsbegriff aufgebaut wird. Diese neutrale Verwendung von „Diskriminierung" gibt es zwar auch im Deutschen, doch wenn es um die Diskriminierung in sozialen Kontexten geht, handelt es sich für gewöhnlich um einen normativen Begriff. Neutral verwende ich stattdessen „Differenzierung" und „differenzierende Ungleichbehandlung". „Diskriminierung" verwende ich in diesem Text im Sinne einer moralisch problematischen Ungleichbehandlung.

In 2.1 werden die sachlichen und in 2.2 die normativen Gründe für das Vorliegen einer Diskriminierung näher bestimmt. 2.3 erläutert kurz statistische Diskriminierung, da diese die Grundlage der Diskriminierung durch maschinelles Lernen ist.

2.1 Sachliche Unangemessenheit

Diskriminierungen liegen dann vor, wenn eine Ungleichbehandlung nicht sachlich angemessen ist. Eine Ungleichbehandlung ist sachlich nicht angemessen, wenn die Art der Ungleichbehandlung nicht dazu geeignet ist, dem Zweck der Ungleichbehandlung zu dienen. Wenn es etwa darum geht, die Person zu finden, die am besten für einen IT-Arbeitsplatz geeignet ist, ist es nicht sachdienlich, Frauen oder Schwarze von vornherein nicht für diesen Arbeitsplatz in Erwägung zu ziehen, da weder Geschlecht noch Hautfarbe eine Qualifikation für den Arbeitsplatz darstellen. In diesem Fall spreche ich immer von einer Diskriminierung. Sachliche Unangemessenheit ist dabei ein hinreichendes, aber kein notwendiges Merkmal der Diskriminierung, da eine sachlich angemessene Ungleichbehandlung aus normativen Gründen diskriminierend sein kann (Britz 2008, 127–130; Avraham 2018, 341f.).

Liegt ein statistischer Zusammenhang zwischen dem Zweck der Ungleichbehandlung und dem Merkmal der Ungleichbehandlung vor (siehe 2.3), ist es möglich, dass die Ungleichbehandlung sachlich angemessen ist. Hier spielen jedoch normative Überlegungen eine erhebliche Rolle dafür, wie stark der Zusammenhang sein muss, damit eine Ungleichbehandlung als sachlich angemessen gelten kann. Umfassend sachlich angemessen ist nur die Betrachtung des Einzelfalls (Britz 2008). Da dies jedoch oft nicht möglich ist, kann auch eine auf Statistik beruhende Ungleichbehandlung sachlich angemessen sein.

Diskriminierung als sachlich unangemessene Ungleichbehandlung entspricht in etwa der „irrelevance discrimination“, wie sie Lippert-Rasmussen (2014, 23) definiert. Diskriminierung liegt nach dieser Theorie dann vor, wenn eine Ungleichbehandlung aufgrund von Gründen erfolgt, die für die Handlungssituation irrelevant sind. Lippert-Rasmussen lehnt diese Theorie ab, da sie weder hinreichende noch notwendige Gründe für das Vorliegen einer Diskriminierung liefern könne. Wie bereits erwähnt, bin ich auch der Meinung, dass eine sachlich unangemessene Ungleichbehandlung nicht notwendig für Diskriminierung ist. Hinreichend ist Lippert-Rasmussen zufolge die sachliche Unangemessenheit deshalb nicht, weil Diskriminierung einen Gruppenbezug hat und das Argument der sachlichen Unangemessenheit diesen nicht automatisch aufweist. In 3.1 argumentiere ich, dass der Gruppenbezug nicht notwendig für Diskriminierung ist. Der zentrale Punkt, für den in diesem Text argumentiert wird – die Möglichkeit der intransparenten Diskriminierung anhand neuer Merkmale durch maschinelles Lernen –, ist jedoch kompatibel mit einem gruppenbasierten Diskriminierungsverständnis.

2.2 Normative Aspekte

Wie bereits erwähnt, verwende ich den Diskriminierungsbegriff so, dass er nicht einfach differenzierend ist, sondern auch eine normative Komponente aufweist. Es geht also um eine moralisch problematische Ungleichbehandlung. Es muss jedoch spezifiziert werden, in welchem Sinn die Ungleichbehandlung moralisch problematisch ist. Mit Lippert-Rasmussen (2014, 29) ist zwischen einer „Pro-tanto“- und einer „All-things-considered“-Version einer moralisch problematischen Ungleichbehandlung zu unterscheiden. „All things considered“ soll hier heißen, dass neben der konkret vorliegenden Ungleichbehandlung auch alle anderen moralischen Aspekte der Situation betrachtet und abgewogen werden. So kann die Ungleichbehandlung für sich genommen problematisch sein, aber moralisch dadurch aufgewogen werden, dass andere Güter damit verwirklicht werden. Ist etwa ein Krankenversicherungssystem ökonomisch nur dann tragbar, wenn geschlechtsbasierte Ungleichbehandlungen stattfinden, könnte die Ungleichbehandlung „all things considered“ moralisch angemessen sein. Wenn man Diskriminierung auf der Grundlage einer solchen Theorie bestimmen wollte, läge hier eine Ungleichbehandlung, aber keine Diskriminierung vor. Diskriminierung kann jedoch genauso wie Körperverletzungen, Diebstahl, Lügen und dergleichen mit anderen moralischen Gütern abgewogen werden, ohne selbst aufzuhören eine Diskriminierung zu sein. Weiterhin ist Lippert-Rasmussen

recht zu geben, dass es immer theoretisch denkbare Fälle gibt, in denen eine Ungleichbehandlung in Anbetracht der Gesamtsituation moralisch gerechtfertigt sein kann – man denke nur an die Verhinderung des Weltuntergangs durch eine Diskriminierung am Arbeitsplatz. Der Diskriminierungsbegriff einer „All-things-considered"-Theorie wäre in der Praxis kaum sinnvoll anzuwenden.

Doch auch die „Pro-tanto"-Version einer moralisch problematischen Ungleichbehandlung als Grundlage des Diskriminierungsbegriffs überzeugt Lippert-Rasmussen (2014, 25) nicht. Hier ist es nun jedoch zunächst wichtig zu bestimmen, wann eine „pro tanto" unmoralische Ungleichbehandlung vorliegt, um die Theorie beurteilen zu können. Im Unterschied zur „All-things-considered"-Version werden hier nur die sachbezogenen normativen Aspekte in den Blick genommen. Eine sachbezogen normativ unangemessene Ungleichbehandlung liegt dann vor, wenn, unabhängig von der sachlichen Angemessenheit, aus normativen Gründen nicht an das Merkmal der Ungleichbehandlung angeknüpft werden sollte. So mag es zwar aus statistischen Gründen sachlich angemessen sein, Frauen aufgrund durchschnittlich höherer verursachter Kosten einen höheren Krankenkassenbeitrag zahlen zu lassen, aus normativen Gründen könnten wir die zusätzliche finanzielle Belastung von Frauen gegenüber Männern im Gesundheitsbereich jedoch für unangemessen und damit diskriminierend halten (eine Frage, die rechtlich in Deutschland noch umstritten ist). Sachbezogen ist diese normative Unangemessenheit, weil sie nur auf den konkret vorliegenden Sachverhalt schaut und nicht „all things considered" bewertet. Auf den konkreten Sachverhalt zu schauen soll hier heißen, bei den normativen Überlegungen nur darauf abzustellen, ob der Person oder Personengruppe, die ungleich behandelt wird, ein unangemessener Schaden entsteht (oder typischerweise entstehen könnte). Die Folgen für andere Personen werden dabei nicht berücksichtigt. Diese können dann bei der Frage eine Rolle spielen, ob die Diskriminierung – „all things considered" – gerechtfertigt werden kann. Was ein unangemessener Schaden ist, hängt von weiteren normativen Überlegungen ab. Hierbei geht es darum, welche Nachteile aufgrund einer sachlich angemessenen Ungleichbehandlung einer Person oder Personengruppe grundsätzlich normativ zuzumuten sind (Britz 2008, 127–130). Dies läuft auch darauf hinaus, den Diskriminierungsschutzzweck in Bezug auf die konkrete Ungleichbehandlung zu konkretisieren, um die Frage der Zumutbarkeit zu klären. Was der Diskriminierungsschutzzweck im Einzelnen ist, ist umstritten und kann und muss an dieser Stelle auch nicht endgültig beant-

wortet werden. In Frage kommen aber etwa Gleichheit, Personalitätsschutz, Kompensation von Nachteilen, Ermächtigung, Verhinderung von Stereotypisierung oder konstellationsspezifische Aspekte – um nur die Punkte zu nennen, die in Britz (2008) Erwähnung finden. Gerade in Auseinandersetzung mit indirekter Diskriminierung wird die Debatte noch komplizierter (Khaitan 2015). Die Abgrenzung zwischen einer gerechtfertigten Diskriminierung und einer nicht diskriminierenden gerechtfertigten Ungleichbehandlung kann in einigen Einzelfällen erhebliche Schwierigkeiten bereiten. Die Abgrenzung und Rechtfertigungen sind jedoch stark kontextabhängig, weshalb an dieser Stelle keine allgemeine Lösung angegeben werden kann.

Lippert-Rasmussen (2014, 25f.) lehnt die moralische „Pro-tanto"-Theorie der Diskriminierung aus drei Gründen ab, die meiner Meinung nach jedoch nicht überzeugend sind.

Erstens würden damit auch unmoralische idiosynkratische Ungleichbehandlungen wie Nepotismus oder die Benachteiligung aufgrund von beliebigen Präferenzen (etwa die Benachteiligung von Leuten aus Omaha, insbesondere solchen, die gut im Hochschulsport waren und die in einem Gebiet leben, dessen Name mindestens ein „s" beinhaltet) als Diskriminierung erfasst werden. In Teil 3.1 argumentiere ich im Anschluss an Thomsen (2013), dass dies dennoch ein plausibles Verständnis von Diskriminierung ist, lege mich auf einen solchen Begriff jedoch nicht fest, da ich nur dafür argumentieren möchte, dass Diskriminierungen anhand neuer Merkmale möglich sind. Neue Merkmale können dabei idiosynkratisch sein oder aber auch Gruppenmerkmale sein. Daher würde ich ebenfalls Lippert-Rasmussens (2014, 29) – von ihm selbst abgelehnte – pro tanto unmoralische Gruppendiskriminierungstheorie akzeptieren, die zusätzlich fordert, dass es sich um benachteiligte Gruppen und nicht um „nur" idiosynkratische Ungleichbehandlungen handelt.

Zweitens sei ein moralischer Diskriminierungsbegriff unplausibel, da sich manche die Frage (sinnvoll) stellen würden, ob eine Diskriminierung moralisch gerechtfertigt ist (Lippert-Rasmussen 2014, 25). Diese Frage macht sicherlich dann Sinn, wenn man von einem neutralen Diskriminierungsbegriff ausgeht. In den deutschsprachigen Debatten wird jedoch für gewöhnlich von einem moralischen Diskriminierungsbegriff ausgegangen. Daher ist hier sicherlich auch die Frage weniger sinnvoll, ob eine Diskriminierung moralisch problematisch ist oder nicht. Weiterhin lässt sich Lippert-Rasmussens Intuition durch seine eigenen Ausführungen weiter entkräften. Wenn man zwischen einer „All-things-considered"- und einer „Pro-tanto"-Version un-

terscheidet, dann zeigt sich, dass in der „Pro-tanto"-Version die Frage nach der moralischen Rechtfertigung einer Diskriminierung durchaus noch sinnvoll gestellt werden kann. Wenn andere moralische Güter in einer Abwägung den Schaden, der durch die Diskriminierung entsteht, aufwiegen, dann kann die Diskriminierung unter Anbetracht aller Tatsachen gerechtfertigt sein, auch wenn die Diskriminierung selbst moralisch problematisch bleibt.

Drittens argumentiert Lippert-Rasmussen (2014, 30), dass in der von mir favorisierten „Pro-tanto"-Version Affirmative Action als Diskriminierung verstanden werden muss. Dies halte ich jedoch nicht nur für unproblematisch, sondern sogar für wünschenswert. Schließlich wird in diesen Fällen nicht umsonst häufig von positiver Diskriminierung gesprochen. Diese Diskriminierungen können dann „all things considered" gerechtfertigt sein, um Nachteile, etwa durch vergangene Diskriminierungen, auszugleichen.

Damit sprechen alle drei Einwände von Lippert-Rasmussen nicht zwingend gegen einen „pro tanto" moralischen (optional gruppenbasierten) Diskriminierungsbegriff. Dieser Diskriminierungsbegriff ist die Grundlage der weiteren Überlegungen und erlaubt es, sachliche und sachbezogen normative Unangemessenheit als entscheidend für das Vorliegen von Diskriminierungen zu verhandeln.

2.3 Statistische Diskriminierung

Diskriminierung durch maschinelles Lernen kann als ein spezieller Fall statistischer Diskriminierung verstanden werden. Daher wird es in diesem Teil darum gehen, statistische Diskriminierung genauer darzustellen und aufzuzeigen, an welchen Stellen Diskriminierung durch maschinelles Lernen Besonderheiten aufweist.[2]

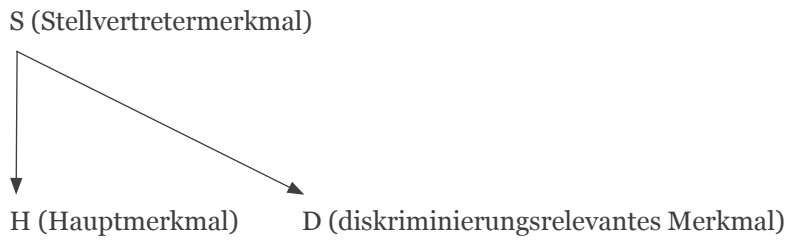
Statistische Unterscheidungen anhand von Stellvertretermerkmalen werden insbesondere dann vorgenommen, wenn das gesuchte Hauptmerkmal nicht direkt festgestellt werden kann. Ob eine Person eine leistungsstarke Arbeitskraft sein wird, lässt sich in der Gegenwart nicht direkt erfassen. Ein Mensch kann versuchen, aufgrund von Erfahrungswissen eine Person diesbezüglich einzuschätzen. Dieses Erfahrungswissen aufzubauen und abzurufen ist aufgrund des benötigten Personals jedoch sehr kostspielig. Weiterhin sind die Einschätzungen aufgrund von Erfahrungswissen oft fehlerhaft und nicht frei von Diskriminierungen. Letztlich ist auch das Erfahrungswissen

2 Die Ausführungen zu statistischer Diskriminierung haben als Grundlage Britz (2008).

eine, wenn auch sehr subjektiv getrübte, statistische Einschätzung. Mit dem Zugang zu großen Datenmengen und der Möglichkeit einer statistischen Auswertung lassen sich (zumindest vermeintlich) objektivere Kriterien mit großer Effizienz gewinnen. So wäre es möglich, dass in einem Unternehmen ein signifikanter Zusammenhang zwischen Geschlecht und Leistungsfähigkeit in einem bestimmten Beruf festgestellt wird. In diesem Fall wäre das gesuchte Hauptmerkmal Leistungsfähigkeit und das Stellvertretermerkmal, mit dem versucht wird das Hauptmerkmal zu erfassen, Geschlecht. Würde ein sexistischer Arbeitgeber direkt nach nur männlichen Arbeitnehmern suchen, läge ein Fall von nichtstatistischer Diskriminierung vor. In dem hier beschriebenen Fall wird erst anhand einer statistisch angenommenen Korrelation anhand des Stellvertretermerkmals Geschlecht diskriminiert.

Findet anhand eines scheinbar neutralen Stellvertretermerkmals eine Ungleichbehandlung statt, die jedoch eine Ungleichbehandlung entlang eines Diskriminierungsmerkmals zur Folge hat, lässt sich von einer indirekten statistischen Diskriminierung sprechen. So könnte ein signifikanter Zusammenhang zwischen Bildungsabschluss und Ausfallrisiko eines Kredites festgestellt werden. Sofern Bildungsabschluss im Kontext der Kreditvergabe als sachlich angemessen und sachbezogen normativ angemessen betrachtet wird und damit nicht als diskriminierungsrelevante Kategorie verstanden wird, läge hier keine Diskriminierung vor. Nun könnte es jedoch sein, dass das Merkmal Herkunft in einem signifikanten Zusammenhang mit Bildungsabschluss steht. Dies würde dazu führen, dass Menschen mit einer bestimmten Herkunft systematisch bei der Kreditvergabe benachteiligt werden. Somit läge eine indirekte Diskriminierung vor, die über das scheinbar neutrale Merkmal Bildungsabschluss vermittelt wäre.[3]

3 Hierbei ist es möglich, dass das Diskriminierungsmerkmal (hier Herkunft) sogar auch signifikant mit dem Hauptmerkmal zusammenhängt. Dann ist es möglich, über eine indirekte statistische Diskriminierung eine verdeckte direkte statistische Diskriminierung durchzuführen. So dürfen etwa KFZ-Versicherungen nicht nach dem verpönten Stellvertretermerkmal Herkunft ausdifferenziert werden, um das Hauptmerkmal Unfallwahrscheinlichkeit zu erfassen. Dies wurde versucht zu umgehen, indem anhand des Autokennzeichens eine Ausdifferenzierung vorgenommen wurde, mit dem eigentlichen Ziel, anhand von Herkunft zu unterscheiden. Da das Stellvertretermerkmal Autokennzeichen signifikant mit Herkunft zusammenhängt, liegt scheinbar eine mittelbare statistische Diskriminierung vor. Da das Stellvertretermerkmal jedoch absichtlich nur dazu dient, anhand von Herkunft zu unterscheiden, liegt in Wirklichkeit eine verdeckte unmittelbare statistische Diskriminierung



Beim maschinellen Lernen geht es zumeist darum, dass ein Algorithmus erlernen soll, geeignete Stellvertretermerkmale zu finden. Wenn es etwa darum geht, die Rückfallwahrscheinlichkeit von Straftäter_innen zu ermitteln (Hauptmerkmal), kann ein lernender Algorithmus anhand von historischen Daten trainiert werden, um Prognosen zu erstellen[4]. Je nachdem welche Daten zur Verfügung stehen, könnte es passieren, dass der Algorithmus anhand von Merkmalen wie Herkunft oder Geschlecht (Stellvertretermerkmale) Prognosen erstellt. Doch es könnte auch passieren, dass der Algorithmus lernt, dass Stellvertretermerkmale besonders wichtig für die Prognose sind, die bisher von Menschen nicht in Erwägung gezogen worden sind und die historisch keine Rolle in Diskriminierungskontexten gespielt haben. Diese Merkmale könnten, wie in der Einleitung erwähnt, unbekannt und unverständlich sein. Außerdem könnte eine Erklärung dafür fehlen, weshalb anhand dieser Merkmale eine gute Prognose über Rückfallwahrscheinlichkeiten erstellt werden kann. Hier liegt es nahe sich zu fragen, ob wir es zulassen wollen, dass ein intransparenter Algorithmus Prognosen erstellt, die relevant dafür sein können, ob jemand aus der Haft entlassen wird.

Ein großes Problem statistisch begründeter Ungleichbehandlungen ist die Einzelfallgerechtigkeit. „Der Betroffene muss sich allein wegen dieses einen (Stellvertreter-)Merkmals einer Regel beugen, die auf zahlreiche Personen, die dieses Merkmal aufweisen, passen mag, die aber seinen Fall nicht richtig erfasst, und darum eigentlich nicht zur Anwendung kommen dürfte“

vor (Britz 2008, 56). Ähnlich könnte natürlich auch eine nichtstatistische Diskriminierung verdeckt werden. So könnte ein rassistisches Versicherungsunternehmen nach Autokennzeichen differenzieren, um zu versuchen, den eigenen Rassismus zu verdecken oder rechtskonform auszugestalten.

4 Es soll mit diesem Beispiel nicht nahegelegt werden, dass Strafe, Strafvollzug und die Ermittlung von Rückfallwahrscheinlichkeiten ein angemessener Umgang mit gesellschaftlichen Problemen sei.

(Britz 2008, 12). Wie schwerwiegend dieses Problem ist, hängt (auch) von der Prognoseleistung der Statistik bzw. des Algorithmus ab.

Anhand von geschlechtlichen biologischen Merkmalen unterschiedliche medizinische Behandlungen durchzuführen kann sachlich angemessen sein (etwa bei statistischem Zusammenhang zwischen diesen Merkmalen und dem Behandlungserfolg), anhand derselben Merkmale aktives Wahlrecht zuzusprechen ist sachlich unangemessen. Die sachliche Angemessenheit hängt in statistischen Zusammenhängen nicht nur von der Prognoseleistung ab. Die Verwendung von Korrelationen, die keine tatsächlich vorliegenden Kausalzusammenhänge abbilden, kann trotz guter Prognoseleistung diskriminierend sein (Avraham 2018, 341f.). Dies wird später in Teil 4 genauer besprochen, wenn es darum geht, ob Intransparenz Wissen über Kausalzusammenhänge und Diskriminierungen verhindert.

Im Zusammenhang mit (statistischer) Diskriminierung geht es nicht nur um Einzelfallgerechtigkeit (Britz 2008, 127–130). Neben der sachlichen Angemessenheit (hier wesentlich statistische Methoden, Prognosen und Kausalzusammenhänge) sind auch normative Vorstellungen darüber, welche Merkmale eine Rolle in dem vorliegenden Sachverhalt spielen sollten (sachbezogen normative Angemessenheit), relevant. Auch bei „richtiger", sachlich angemessener Differenzierung oder zutreffender Prognose kann eine Diskriminierung vorliegen, etwa dann, wenn unangemessen anhand sachbezogen normativ relevanter Merkmale wie Herkunft oder Geschlecht ungleich behandelt wird. Geschlecht mag statistisch ein guter Indikator für zukünftig verursachte Krankenkosten sein, der sogar kausale Prozesse abbilden mag, den wir aus normativen Gründen aber dennoch als unangemessen und damit diskriminierend für die Berechnung von gesetzlichen Krankenversicherungsbeiträgen ansehen könnten (zum komplizierten Streit um diese Frage etwa Avraham [2018, 337]).

Wenn wir Entscheidungen an intransparente Algorithmen abgeben, nehmen wir uns die Chance, die Entscheidungsgrundlagen kritisch zu reflektieren und zu bewerten. Die sachliche Angemessenheit erschöpft sich – wie bereits gezeigt – nicht in der Verwendung signifikanter Korrelationen und die menschliche Bewertung der normativen Angemessenheit erfordert zumindest eine Kenntnis der statistischen Zusammenhänge und der verwendeten Merkmale. Ein vergleichbares Problem kann generell bei indirekter (statistischer) Diskriminierung entstehen. Auch diese kann, wenn auch in einem anderen Sinne, intransparent sein und muss zum Teil erst mit aufwendiger sozialwissenschaftlicher Forschung offengelegt werden.

In Abschnitt 4 wird näher auf die Ursachen und Folgen intransparenter Diskriminierung im Kontext des maschinellen Lernens eingegangen. Zuvor aber müssen hierfür in Abschnitt 3 letzte begriffliche Klärungen in Bezug auf Diskriminierungsmerkmale, Intentionalität und mentale Eigenschaften vorgenommen werden.

3. Begriffliche Probleme der Diskriminierung durch maschinelles Lernen

Es stellen sich im Zusammenhang mit maschinellem Lernen insbesondere zwei Herausforderungen für den Diskriminierungsbegriff: Erstens stellt sich die Frage, inwieweit die Merkmale, anhand derer der Algorithmus differenziert, als Diskriminierungsmerkmale in Frage kommen, und zweitens muss geklärt werden, inwiefern leblose Maschinen im Vergleich zu Menschen überhaupt dazu in der Lage sind zu diskriminieren.

Maschinelles Lernen differenziert auch anhand von Merkmalen, anhand derer historisch keine Diskriminierung stattgefunden haben muss. In 3.1 wird gezeigt, wann auch anhand dieser neuen Merkmale eine Diskriminierung stattfinden kann. Weiterhin wird in 3.2 argumentiert, dass Diskriminierung nicht von diskriminierenden Absichten, generellen Handlungsabsichten oder mentalen Repräsentationen von Differenzen abhängig ist, weshalb auch Algorithmen, die nicht über mentale Eigenschaften verfügen, diskriminieren können.

3.1 Diskriminierungsmerkmale

In deutschen gesetzlichen Regelungen werden abschließende Listen aufgestellt, die diejenigen Merkmale festlegen, auf deren Grundlage eine Ungleichbehandlung verboten ist bzw. unter besonderem Rechtfertigungsdruck steht (Art 3 GG, §1 AGG). Findet dennoch eine solche Ungleichbehandlung statt, wird diese oft als Diskriminierung bezeichnet. In europarechtlichen Regelungen finden sich zumeist keine abschließenden Listen (Art 14 EMRK, Art 21 GRC). Außerdem werden dort auch deutlich mehr Merkmale explizit genannt. Im deutschen Recht werden die im GG und AGG nicht explizit genannten Merkmale entweder in anderen rechtlichen Regelungen genannt oder vom allgemeinen Gleichbehandlungsgrundsatz des Art 3 Abs. 1 GG abgedeckt. Unterschiede ergeben sich jedoch im unterschiedlich starken Rechtfertigungsbedarf der Ungleichbehandlungen.

Grundsätzlich stellt sich die Frage, ob Diskriminierung nur anhand besonderer Gruppenmerkmale stattfinden kann oder ob auch idiosynkratische Ungleichbehandlungen diskriminierend sein können. Wenn vertreten wird, dass nur bestimmte Gruppenmerkmale Diskriminierung begründen können, muss ein Kriterium angegeben werden, um diese besonderen Gruppenmerkmale identifizieren zu können. Eine schwer zu verteidigende Strategie besteht darin, bestimmte Gruppenmerkmale – wie etwa Geschlecht oder Herkunft – an sich schon als diskriminierungsrelevant zu verstehen. Die vielleicht erfolgversprechendste Strategie dies zu rechtfertigen, besteht in der Annahme, dass unveränderliche Merkmale, für die wir keine Verantwortung tragen, an sich schon diskriminierungsrelevant seien. Thomsen (2013, 129–137) zeigt jedoch, dass dies unplausibel ist: Zu viel wird eingeschlossen und zu viel auch ausgeschlossen. Wir haben nahezu unendlich viele unveränderliche Eigenschaften, für die wir keine Verantwortung haben, doch die meisten davon scheinen nicht an sich schon diskriminierungsrelevant zu sein (etwa die Menge der Vokale in unserem Vornamen). Gleichzeitig ist etwa unsere Religion durchaus willentlich änderbar und wäre damit nicht diskriminierungsrelevant.

Aussichtsreicher scheint es, wenn man nicht versucht, Gruppenmerkmale zu bestimmen, die an sich schon diskriminierungsrelevant sind, sondern kontextabhängige Eigenschaften identifiziert, die Gruppenmerkmale diskriminierungsrelevant werden lassen. Thomsen (2013, 137–143) kritisiert die seiner Meinung nach aussichtsreichsten Ansätze, die versuchen, diese Eigenschaften über einen zusätzlichen Schaden der Ungleichbehandlung zu bestimmen. Die Grundidee ist, dass unter bestimmten soziohistorischen Bedingungen bestimmte Gruppen von Benachteiligungen besonders hart getroffen werden. Hierfür müsste seiner Meinung nach jedoch gezeigt werden, dass es einen Grenzwert gibt, ab dem diese besondere Härte der Benachteiligung vorliegt. Nicht nur sei dieser Grenzwert schwer zu bestimmen, auch sei es unplausibel anzunehmen, dass dieser Grenzwert in idiosynkratischen Benachteiligungen niemals erfüllt sei.

Der prominenteste Ansatz, Gruppenmerkmale zu bestimmen, dürfte von Lippert-Rasmussen (2014) stammen. Er argumentiert, dass nur sozial saliente Gruppen diskriminiert werden können, und definiert soziale Salienz wie folgt: „A group is socially salient if perceived membership of it is important to the structure of social interactions across a wide range of social contexts“ (Lippert-Rasmussen 2014, 30). Zu seiner allgemeinen Definition von Diskriminierung gehört jedoch auch eine davon separate Bedingung, die

diesen Aspekt schon erfasst: „Φ is a relevant type of act, policy, or practice, and there are many acts etc. of this type, and this fact makes people with P (or some subgroup of these people) worse off relative to others, or Φ is a relevant type of act etc., and many acts etc. of this type would make people with P worse off relative to others“ (Lippert-Rasmussen 2014, 28). Wenn etwas ein vielfach auftretendes Verhalten ist, das eine Gruppe schlechterstellt oder stellen würde, dann ist diese Gruppe sicherlich wichtig für soziale Interaktionen in einem weitreichenden sozialen Kontext und damit salient. Hiermit zielt Lippert-Rasmussen im Grunde genommen implizit darauf ab, dass Diskriminierung eine Strukturkategorie sein soll, die keine einzelnen idiosynkratischen Ungleichbehandlungen erfassen soll. Daher wird Nepotismus für ihn erst dann zu einer Diskriminierung, wenn die Bevorzugung von Familienmitgliedern bestimmter Großfamilien umfassend betrieben wird (Lippert-Rasmussen 2014, 34f.).

Eidelson (2015) argumentiert gegen Lippert-Rasmussen, dass Diskriminierung keine salienten sozialen Gruppen voraussetzt. Die Plausibilität des Salienzkriteriums leite sich aus einem alltagssprachlichen moralisierenden Gebrauch von Diskriminierung her. Wäre Nepotismus gesellschaftlich geächtet, würde Nepotismus auch alltagssprachlich als Diskriminierung gelten. Thomsen (2013, 142f.) hingegen gesteht Lippert-Rasmussen zu, zumindest ein pragmatisches Unterscheidungsmerkmal gefunden zu haben, das zwar keinen grundsätzlichen begrifflichen Unterschied zwischen der Benachteiligung von sozial salienten Gruppen und idiosynkratischen Benachteiligungen etablieren kann, aber auf einen graduellen Unterschied in der Härte der Benachteiligung verweisen kann.

Insbesondere wenn es im Zusammenhang mit Diskriminierung um abschließende Listen von Gruppenmerkmalen geht, ist das Ziel für gewöhnlich ausgewiesene Gruppen unter einen besonderen Schutz zu stellen. Die Rechtfertigungshürde für die Ungleichbehandlung besonders geschützter Gruppen ist damit höher als in einem idiosynkratischen Fall. Dies macht vor dem Hintergrund der besonderen Verletzlichkeit derjenigen Gruppen Sinn, die in diesen abschließenden Listen aufgezählt werden. Diese Gruppen dürfen unter besonderen Bedingungen bei entsprechender Rechtfertigung (sachliche und sachbezogen normative Angemessenheit) jedoch auch ungleich behandelt werden, ohne dass es sich um Diskriminierung handelt (etwa Ungleichbehandlung nach Geschlecht bei bestimmten medizinischen Behandlungen). Da es also im Endeffekt immer auf die Rechtfertigung im Einzelfall ankommt, unterscheiden sich Ungleichbehandlungen von ausgewiesenen

Gruppen und idiosynkratische Ungleichbehandlungen nur in ihrem Rechtfertigungsbedarf. Damit lässt sich kein Merkmal ausweisen, anhand dessen – ungeachtet der Rechtfertigung im Einzelfall – eine Ungleichbehandlung immer diskriminierend ist. Abgesehen von praktischen und normativen Gründen, Ungleichbehandlungen anhand bestimmter Merkmale einem höheren Rechtfertigungsdruck auszusetzen und damit besonders sensible Gruppen unter einen effektiveren Schutz zu stellen, lässt sich auch kein Merkmal ausweisen, anhand dessen eine Ungleichbehandlung grundsätzlich nicht diskriminierend sein kann. Diese pragmatischen Gründe können einen stärker eingeschränkten Gebrauch des Diskriminierungsbegriffs rechtfertigen, sprechen aber nicht grundsätzlich dagegen, auch idiosynkratische Ungleichbehandlungen als Diskriminierung zu verstehen.

Damit kommen die Merkmale, anhand derer Algorithmen lernen zu differenzieren, grundsätzlich als Diskriminierungsmerkmale in Frage. Dann muss jedoch geklärt werden, inwiefern die Neuartigkeit der Merkmale im Kontext maschinellen Lernens besondere Herausforderungen nach sich zieht. Auch rein menschliche Benachteiligungen können anhand neuartiger Merkmale stattfinden. Nennenswerte Unterschiede ergeben sich aber in Bezug auf die Identifikation dieser Merkmale und die Systematizität der Benachteiligungen.

Die Identifikation neuer Merkmale erweist sich als schwierig, da nicht bekannt ist, nach welchen Merkmalen genau gesucht werden soll. Wie bereits erwähnt, wurde die Diskriminierung durch maschinelles Lernen anhand von Merkmalen, die explizit unter Diskriminierungsschutz stehen, intensiv diskutiert. Hierbei wird in dem Algorithmus oder in dem Output des Algorithmus nach Differenzierungen gesucht, die eine Benachteiligung oder schlechtere Bewertung von Gruppen nach sich ziehen, die unter Diskriminierungsschutz stehen. Systematische Ungleichbehandlungen lassen sich hierbei in Algorithmen verhältnismäßig gut identifizieren.[5] Dies gilt nicht nur für Merkmale, die bereits in Listen aufgeführt werden, sondern grundsätzlich auch für neue Merkmale, sofern diese schon vor dem Suchvorgang bekannt sind und angegeben werden können. Wenn also pragmatisch ein

5 Ein großes Problem bei dieser Suche tritt auf, wenn indirekte Benachteiligungen (indirekte Diskriminierungen) auftreten. Hierbei kann es sich um indirekte Diskriminierungen anhand von Merkmalen handeln, die in den für den Algorithmus zur Verfügung stehenden Merkmalsdaten nicht vorhanden sind. Diese Formen der Benachteiligung können erst durch zum Teil aufwendige Untersuchungen erkannt werden.

kontextuelles Gruppenmerkmal für die Bestimmung von Diskriminierungen herangezogen wird (etwa durch soziale Salienz), lässt sich der Algorithmus im Hinblick auf Ungleichbehandlungen anhand dieses neuen Merkmals relativ gut untersuchen. Deutlich schwieriger wird es, wenn die Merkmale, die einen besonderen Diskriminierungsschutz rechtfertigen, nicht schon vorher feststehen. Dann muss generell nach Ungleichbehandlungen anhand aller möglichen Merkmale oder Merkmalskombinationen gesucht werden und jede diese Ungleichbehandlungen auf ihren potentiell diskriminierenden Charakter hin untersucht werden. Dies würde den Effizienzgewinn statistischer Methoden (hier des maschinellen Lernens) zunichtemachen. Schließlich müsste man theoretisch alle Fälle einzeln ansehen, um zu überprüfen, ob die Ungleichbehandlungen sachlich und sachbezogen normativ angemessen sind. Zwar können bestimmte Merkmale vorher festgelegt werden, die in keinem Fall diskriminierungsschutzrelevant sind, trotzdem würden noch extrem viele Merkmale und Merkmalskombinationen übrig bleiben, die untersucht werden müssten.

Anhand der bis hierhin verfolgten Debatte um Diskriminierungsmerkmale ist davon auszugehen, dass zumindest systematische (nicht idiosynkratische) Ungleichbehandlungen anhand neuer Merkmale, die nicht in abschließenden Listen aufgeführt werden, möglich sind.[6] Daher möchte ich abschließend zu der Frage der Diskriminierungsmerkmale die Systematizität von Benachteiligungen als Kriterium zur Identifikation von Diskriminierungen näher beleuchten. Systematizität lässt sich im Kontext des maschinellen Lernens auf zumindest zwei unterschiedliche Arten interpretieren: anhand der gruppenbezogenen Effekte oder anhand der Ursachen.

Systematizität unterscheidet sich kaum von sozialer Salienz im Sinne von Lippert-Rasmussen (2014), wenn sie anhand der gruppenbezogenen Effekte bestimmt wird. Werden etwa Menschen mit einem bestimmten Reiseverhalten durch einen Algorithmus als potentielle Terroristen eingestuft und resultieren hieraus umfassende Einschränkungen in vielen Kontexten, könnte dies als eine Diskriminierung anhand des Reiseverhaltens verstanden werden. Der systematische Charakter der Ungleichbehandlung käme hier erst durch den systematischen Einsatz des Algorithmus zustande.[7]

6 Abseits der Diskriminierung durch Algorithmen stellt schon die intersektionale Diskriminierung eine Herausforderung für die Aufstellung von Listen möglicher Diskriminierungsmerkmale dar (Stoljar 2018).

7 Dies wäre jedoch, zumindest nach den Prinzipien der deutschen Rechtspre-

Die zweite Position lässt sich gewinnen, wenn angenommen wird, dass der systematische Charakter der Benachteiligung nicht aufgrund der Effekte auf eine Gruppe, die bestimmte Merkmale teilt, zustande kommt, sondern aufgrund einer gemeinsamen verursachenden Struktur. Die Diskriminierung durch maschinelles Lernen ist dann systematisch, weil ein einzelner Algorithmus systematisch Menschen benachteiligt. Verwendet der Algorithmus eine gewichtete Merkmalskombination aus tausenden Merkmalen, mag jede einzelne Entscheidung kontext- und personenspezifisch sein und damit idiosynkratisch. Die betroffenen Personen verfügen nicht über gemeinsame Merkmale der Ungleichbehandlung, die sie zu einer Gruppe machen würde. Aber die Ungleichbehandlung wird von einer einzelnen Struktur getragen, die über weite Kontexte und auf viele Personen angewandt wird. Damit handelt es sich um systematische Ungleichbehandlungen durch einen Algorithmus, der weitreichend eingesetzt wird. Dies würde einen besonderen Diskriminierungsschutz vor solchen Strukturen sinnvoll erscheinen lassen. Dieser besondere Bedarf eines Diskriminierungsschutzes rechtfertigt es, systematische Ungleichbehandlungen durch einen Algorithmus trotz fehlender gemeinsamer Gruppenmerkmale der ungleich behandelten Personen als Diskriminierung zu verstehen. Eine vielleicht kontraintuitive Konsequenz dieses Ansatzes ist, dass etwa willkürliche Ungleichbehandlungen durch einen Diktator, die keinem speziellen Muster folgen, als Diskriminierung zu verstehen wären, da sie von einer Ursache getragen werden und die betroffenen Personen im Hinblick auf diese Ursache eine Gruppe bilden.

Während systematische Benachteiligungen durch einen Algorithmus (insbesondere in der Ursachenlesart) wenig transparent sein können und trotz ihres systematischen Charakters erst ein gewisser Aufwand nötig sein kann, um diese zu identifizieren, sind systematische Benachteiligungen durch Menschen oft deutlich leichter zu erkennen. Institutionelle Benachteiligungen (und Diskriminierungen) basieren oft auf expliziten Regeln. Zwischenmenschliche Benachteiligungen können zwar unabsichtlich oder unbewusst geschehen, involvieren aber auch häufig Intentionen oder Annahmen, deren diskriminierender Charakter – zumindest bei systematischen Ungleichbehandlungen – leicht erkennbar ist. Die von Benachteiligung betrof-

chung, die allerdings keine Diskriminierung anhand des Reiseverhaltens kennt, nur dann der Fall, wenn die Ungleichbehandlung aufgrund des Reiseverhaltens nicht verhältnismäßig war. Britz (2008, 151ff.) erläutert detaillierter die Frage der Verhältnismäßigkeit bei Benachteiligungen in Staat-Bürger_in- und in Bürger_in-Bürger_in-Verhältnissen.

fenen Personen wissen außerdem nicht selten um diese Benachteiligungen und kennen die Merkmale, an die angeknüpft wird. Daher ist menschliche systematische Benachteiligung relativ transparent. Systematische indirekte menschliche Benachteiligungen können jedoch ähnlich intransparent sein wie algorithmische Benachteiligungen. Schließlich handelt es sich um einen nichtintendierten Effekt. Die Verfahren zur Offenlegung intransparenter indirekter menschlicher Benachteiligung und der Benachteiligung durch maschinelles Lernen unterscheiden sich. Die Effekte menschlicher indirekter Benachteiligung müssen mit sozialwissenschaftlichen empirischen Verfahren identifiziert werden. Die Benachteiligungen durch Algorithmen können zwar auch mit solchen Verfahren identifiziert werden, für gewöhnlich wird jedoch stattdessen der Output des Algorithmus computergestützt analysiert. Die Analyse des Outputs gestaltet sich jedoch bei neuen Merkmalen nicht so einfach. In Abschnitt 4 wird sich nochmals zeigen, dass der intransparente Charakter maschinellen Lernens ein großes Hindernis für diese Offenlegung sein kann.

3.2 Intentionalität und mentale Eigenschaften

Mentale Eigenschaften können in verschiedenen Hinsichten relevant für Diskriminierung sein. Sie können für die Repräsentation der Differenzierung entscheidend sein, für Absichten bei der Ungleichbehandlung oder es kann von ihnen abhängen, wie eine Differenzierung moralisch bewertet wird. Da ein Algorithmus nicht über mentale Eigenschaften verfügt, könnte man behaupten, dass er deshalb auch nicht diskriminierend sein kann. Es bieten sich zumindest drei Möglichkeiten an, mit diesem Problem umzugehen: Erstens kann man versuchen, Eigenschaften des maschinellen Lernens analog zu mentalen Eigenschaften zu deuten (damit könnte der Algorithmus analog als diskriminierend verstanden werden), zweitens kann man versuchen, die relevanten mentalen Eigenschaften stattdessen bei Akteuren zu finden, die in einem explanatorischen Zusammenhang zur Differenzierung oder Ungleichbehandlung des Algorithmus stehen (damit wären menschliche Akteure, die den Algorithmus entwickeln, einsetzen oder auf der Grundlage von Empfehlungen des Algorithmus handeln, diskriminierend) und drittens kann man in Frage stellen, dass mentale Eigenschaften für Diskriminierungen überhaupt relevant seien (damit würde Diskriminierung keine mentalen Eigenschaften voraussetzen und Algorithmen könnten auch ohne diese diskriminierend sein).

Wenig überzeugend ist die Forderung, dass der moralisch problematische Charakter einer Diskriminierung notwendig auf mentalen Eigenschaften basieren muss. Im Rahmen einer tugendethischen Betrachtung, wie sie Garcia (2018, 178) anstellt, sind die mentalen Eigenschaften der diskriminierenden Person essentiell für die Frage, ob eine Diskriminierung vorliegt. Dies ist insofern intuitiv plausibel, als dass wir oft die Geisteshaltung der diskriminierenden Person für besonders kritikwürdig halten. Indirekte Diskriminierung lässt sich jedoch hierdurch kaum mehr erfassen, da es bei dieser gerade nicht auf die Geisteshaltung ankommt, sondern auf die davon unabhängigen Handlungsfolgen. Konsequenterweise müsste indirekte Diskriminierung nicht als Diskriminierung gelten, wie es etwa auch bei dem respektbasierten Ansatz von Eidelson (2015) der Fall ist. Doch selbst die direkte Diskriminierung ist damit nicht überzeugend erfasst. So besteht der Diskriminierungsschutzzweck wesentlich darin, unsachgemäße Ungleichbehandlung, Persönlichkeitsschutz und nachteilhafte Folgen zu verhindern (Britz 2008, 138–211). Schadens- oder effektbasierte Ansätze (etwa Lippert-Rasmussen 2014) sind hier überzeugender. Dies bedeutet jedoch nicht, dass zumindest ein Aspekt, den wir an manchen Diskriminierungen besonders wichtig finden können, die problematische Geisteshaltung ist, die die Diskriminierung trägt. Die moralische Verurteilung rassistischer Diskriminierung zielt schließlich nicht nur auf die Ungleichbehandlung ab, sondern schon auf die rassistische Geisteshaltung. Diskriminierung hat jedoch eine Vielzahl von problematischen Aspekten, von denen die diskriminierende Geisteshaltung nur ein Aspekt ist, der jedoch nicht notwendig vorhanden sein muss. Gleichzeitig erleichtert das Vorliegen einer verurteilenswerten Geisteshaltung auch die Zurechenbarkeit von Verantwortung und das Aufzeigen von Änderungsmöglichkeiten. Bei strukturellen Diskriminierungen, aber auch bei Diskriminierungen durch maschinelles Lernen, ist nicht unmittelbar klar, was genau geändert werden muss und wer Adressat dieser Änderungen ist.

Diskriminierung wird zumeist als eine Handlung, als eine Ungleichbehandlung verstanden. Handlungen setzen Intentionalität voraus und z.T. wird auch aus der Intentionalität der moralisch problematische Charakter der Ungleichbehandlung abgeleitet. Lippert-Rasmussen (2014, 78) argumentiert, dass Strukturen wegen ihrer Handlungsunfähigkeit nur dazu in der Lage seien, indirekte Diskriminierungen zu vollziehen. Garcia (2018, 176) behauptet, dass beispielsweise Strukturen nicht handeln können und nur insofern diskriminierend sein können, als dass sie mit den diskriminierenden Einstellungen von handelnden Akteuren „infiziert“ werden. Ähnliches ließe

sich sicherlich über lernende Algorithmen behaupten. Diese erlernen ihre diskriminierenden Unterscheidungen und Verhaltensweisen oft an vorbelasteten Trainingsdaten. Diese Trainingsdaten können eine diskriminierende Gesellschaft reflektieren, die so geworden ist, wie sie ist, weil Menschen mit diskriminierenden Einstellungen in ihr gehandelt haben. Dieser Umweg ist jedoch groß, nicht immer leicht festzustellen und auch nicht für alle Formen der Diskriminierung durch maschinelles Lernen geeignet. So kann maschinelles Lernen auch neue Formen der Diskriminierung in die Welt bringen. Hier ist unklar, wie der Algorithmus mit diesen „infiziert" worden sein soll. Weiterhin ist es nicht überzeugend, dass der Diskriminierungsschutzzweck sich ausschließlich auf Handlungen erstrecken sollte. Strukturen oder lernende Algorithmen, die Auswirkungen haben, die Personen ungleich nachteilhaft treffen, sollten zumindest von effekt- oder schadensbasierten Ansätzen mit abgedeckt werden.[8]

Die differenzierende Behandlung als Spezifikum einer Diskriminierung lässt sich dabei auch ohne Verweis auf mentales Vokabular verstehen. Bewertungen, Einstufungen oder automatisierte Entscheidungen durch Algorithmen sind in ihren Auswirkungen und ihrem differenzierenden Charakter Handlungen im Kontext von Diskriminierungen hinreichend ähnlich, sodass diese auch bei fehlenden mentalen Eigenschaften, die das Verhalten eines Algorithmus zu einer Handlung machen würden, vom Diskriminierungsschutz erfasst werden sollten.

Wenn wir für die Definition von Diskriminierung weder auf eine moralisch verwerfliche Geisteshaltung noch auf Handlungen im engen Sinn verweisen müssen, so ließe sich zumindest noch behaupten, dass Algorithmen keine Überzeugungen über unterschiedliche Merkmale haben können und damit auch nicht auf Grundlage dieser Überzeugungen ungleich behandeln können. Schon bei menschlicher Diskriminierung werden die Anforderungen an die bewusste Anknüpfung an Merkmale insofern deutlich geschwächt, als dass auch unbewusste Vorurteile und Kategorisierungen als hinreichend angesehen werden. So wird Altman (2011) vielfach dafür kritisiert, intentio-

8 Auch wenn ich nicht glaube, dass Algorithmen handeln können, rede ich oft darüber, dass Algorithmen eine Ungleichbehandlung vornehmen können. Letztlich geht es hierbei nur darum, dass der Algorithmus einen ungleichen Effekt auf Personen oder Personengruppen hat. Insofern angenommen werden sollte, dass „Behandlung" ein Handeln impliziert, möchte ich den Ausdruck hier eher metaphorisch verstanden wissen.

nale Ungleichbehandlung als Bedingung für direkte Diskriminierung zu fordern (Eidelson 2015, 23, Lippert-Rasmussen 2014, 56–58). Altman (2016) ist dieser Kritik auch gefolgt und hat die Anforderung abgeschwächt. Menschen müssen sich ihrer differenzierenden Betrachtungsweise nicht bewusst sein und können dennoch aufgrund einer vorbewussten Differenzierung ungleich behandeln (Lippert-Rasmussen 2014, 36ff.). Eine konsequente Fortführung des Gedankens besteht darin, dass keine bewusste Differenz nötig und auch kein Bewusstsein nötig ist, um von „Diskriminierung" sprechen zu können. Die vorgenommene Differenzierung muss nur ursächlich für eine Ungleichbehandlung sein. Wenn die Differenzierung nicht mental sein muss, erlaubt dies nicht nur Algorithmen als diskriminierend zu verstehen, sondern auch Gesetze, ohne dabei wie Lippert-Rasmussen (2014, 41) auf den mentalen Gehalt der Gesetzgeber zu verweisen.

Khaitan (2015) definiert Diskriminierung ohne expliziten Bezug zu mentalen Eigenschaften oder der Repräsentation von Differenz. Motiviert ist dies durch die Schwierigkeit, zwischen direkter und indirekter Diskriminierung zu unterscheiden. Wenig überzeugend wird diese Unterscheidung oft anhand von Intentionalität gemacht (Altman 2011). Wenn die Diskriminierung nicht beabsichtigt oder zumindest nicht fahrlässig war, handelt es sich demnach um indirekte Diskriminierung. Die Abgrenzung ist dennoch nicht immer klar zu vollziehen. Insbesondere an die britische Rechtsprechung anschließend schlägt Khaitan (2015) stattdessen vor, Diskriminierung nicht von mentalen Eigenschaften der diskriminierenden Person abhängig zu machen, sondern alleine von benachteiligenden Effekten, die in einer Korrelation zur Diskriminierungskategorie stehen (Khaitan 2015, 166). Der Unterschied zwischen direkter Diskriminierung und indirekter Diskriminierung ist hierbei graduell. Direkt ist die Diskriminierung, wenn eine vollständige Deckung zwischen der benachteiligten Personengruppe und der Gruppe mit dem Diskriminierungsmerkmal vorliegt, und indirekt, wenn nur eine teilweise Deckungsgleichheit dieser Gruppen vorliegt. Bei einer solchen rein effektbasierten Definition von Diskriminierung stellen die fehlenden mentalen Eigenschaften eines Algorithmus keinen Hindernisgrund dafür dar, diesen als diskriminierend zu klassifizieren. Lippert-Rasmussen (2016) weist jedoch auf den schon von Khaitan (2015) bemerkten kontraintuitiven Charakter dieser Definition hin und zeigt an einigen Gegenbeispielen, dass diese Definition wenig überzeugend ist. Wie bereits diskutiert, erfordert Diskriminierung jedoch auch ohne die Verwendung dieser Definition keinen Verweis auf mentale Eigenschaften.

In Abschnitt 3 hat sich gezeigt, dass grundsätzlich jedes Merkmal als Diskriminierungsmerkmal in Frage kommt und Intentionalität und mentale Eigenschaften keine Voraussetzung für Diskriminierung sind. Damit ist der Weg geebnet, um im folgenden Abschnitt die Diskriminierung durch Algorithmen anhand neuartiger Merkmale zu untersuchen.

4 Benachteiligung anhand intransparenter Merkmale und Verfahren

Wie in der Einleitung erwähnt, besteht ein besonderes Problem maschinellen Lernens darin, dass die verwendeten Merkmale und Verfahren nicht immer transparent sind. Dem wird in der Forschung seit Längerem versucht zu begegnen, indem eine interpretierbare Künstliche Intelligenz bzw. „explainable artificial intelligence“ (XAI) entwickelt wird. So wird auch im Erwägungsgrund 71 zur europäischen Datenschutzgrundverordnung (DSGVO) eine Erklärung für die Gründe der automatisierten Entscheidungen gefordert. Dieser explizite Bezug auf eine Erklärung wurde jedoch beim endgültigen Gesetzestext nicht mehr berücksichtigt und fehlt daher auch in dem entsprechenden Artikel 22 der DSGVO. In Frankreich hingegen wird der Anspruch auf eine Erklärung im „Loi pour une République numérique“ sogar konkretisiert. Auch in den USA existiert zumindest in Bezug auf Kreditscoring ein solches Recht. Was eine solche Erklärung leisten soll, ist jedoch strittig.

In den Fällen, in denen wir eine Erklärung für maschinelle Entscheidungen benötigen, aber aufgrund der Intransparenz des Algorithmus über keine verfügen, stehen uns oft nur Post-hoc-Interpretationen (Lipton 2017) zur Verfügung. Diese sagen uns zwar nicht genau, was im Algorithmus tatsächlich passiert, aber können uns verstehbare Annährungen geben. Dies kann nach Lipton (2017) etwa über die Generierung alltagssprachlicher Erklärungen, Visualisierungen, lokaler oder beispielhafter Erklärungen geschehen. Da diese jedoch nicht die tatsächlichen Lernprozesse und Funktionsweisen des Algorithmus darstellen, liefern sie nur mögliche Erklärungen für Ungleichbehandlungen, die der tatsächlichen Erklärung jedoch ähnlich sein könnten. Dabei vereinfachen diese Erklärungen komplexe Zusammenhänge. Wird etwa in einer solchen Erklärung nur das einflussreichste Merkmal für eine Klassifikation hervorgehoben (z. B. Einkommen beim Kreditscoring), bleiben möglicherweise diskriminierende Merkmale mit einem geringeren Einfluss verborgen. Ähnlich verhält es sich, wenn etwa bei der Auswertung

von Bildern der Bereich des Bildes markiert wird, der einen großen Einfluss auf die Klassifikation hatte. Beinhalten diese Bildbereiche sensible Merkmale, anhand derer Rückschlüsse auf Geschlecht oder Hautfarbe gezogen werden können, so ist dies ein Indiz für Diskriminierung. Zur Verifizierung, ob tatsächlich eine Diskriminierung vorliegt, werden jedoch genauere Erklärungen benötigt.

Zumindest drei unterschiedliche Hinsichten der Intransparenz sind zu unterscheiden, die für den Kontext der Diskriminierung relevant sind und die die Erklärbarkeit algorithmenbasierter Entscheidungen einschränken. Erstens kann es unbekannt sein, welche Merkmale für eine Klassifikation, Bewertung oder Entscheidung herangezogen werden (4.1). Zweitens ist es möglich, dass die verwendeten Merkmale, selbst wenn sie bekannt sind, nicht intuitiv verstehbar sind, etwa wenn sie hochkomplex sind oder keine alltagssprachliche Beschreibung dieser Merkmale zur Verfügung steht (4.2). Und drittens ist nicht unbedingt klar, warum und wie diese Merkmale verwendet werden (4.3).

4.1 Unbekannte Merkmale

Vielleicht wissen wir, dass der Algorithmus gut vorhersagen kann, bei welchen Personen etwa ein hohes Kreditausfallrisiko besteht. Das heißt aber nicht, dass wir wissen, anhand welcher Merkmale der Algorithmus diese Vorhersage getroffen hat. Damit ist auch nicht unmittelbar bekannt, ob eine Diskriminierung stattfindet. Dieses Problem ist bis zu einem gewissen Grad mit menschlicher Diskriminierung vergleichbar. Anders als bei expliziten Regeln oder Gesetzen sind auch die Gründe oder Ursachen menschlicher Ungleichbehandlung nicht immer transparent. Die Auskunft der Person, die die Ungleichbehandlung vornimmt, kann unwahr sein. Manchmal ist dieser Person sogar selbst nicht einmal transparent, warum genau sie die Ungleichbehandlung vornimmt (siehe 3.2). Hier sind wir auf Indizien oder die systematische Untersuchung des Verhaltens angewiesen. Dies ist vergleichbar mit einem Algorithmus, der keine direkte Auskunft darüber geben kann, anhand welcher Merkmale er ungleich behandelt. Hier gibt es verschiedene Möglichkeiten, das Verhalten des Algorithmus zu untersuchen. Wenn es sich um einen leicht interpretierbaren Algorithmus handelt (etwa in Fällen einer einfachen linearen Regression oder eines einfachen Entscheidungsbaums) und dieser offengelegt wird, können wir nachschauen, welche Merkmale für die Ungleichbehandlungen entscheidend waren. In allen anderen Fällen müssen wir jedoch – ähnlich wie bei Menschen – die getroffenen Entschei-

dungen erst systematisch untersuchen. Hierzu wird anhand der Input/Output-Relationen (also der Relationen zwischen den durch Eingabedaten dargestellten Entscheidungsfällen und den durch Ausgabedaten dargestellten Entscheidungen als Ergebnis des Berechnungsverfahrens) statistisch überprüft, welchen Anteil diskriminierungsrelevante Merkmale für den Output haben.[9] Wird nicht mit abschließenden Listen für diskriminierungsrelevante Merkmale gearbeitet, kann dies einen erheblichen Arbeitsaufwand bedeuten, der insofern erschwert wird und unklar ist, als dass nicht klar ist, wonach überhaupt genau gesucht werden soll.

Die Untersuchung von Input/Output-Relationen im Hinblick auf diskriminierungsrelevante Merkmale ist jedoch nur dann eine erfolgversprechende Strategie, wenn die Merkmale, anhand derer differenziert wird, auch verständlich sind. Dies ist jedoch, wie wir im nächsten Abschnitt sehen werden, nicht immer der Fall.

4.2 Fehlende Verständlichkeit

Lipton (2017) unterscheidet drei andere als die in diesem Abschnitt 4 genannten Aspekte, bezüglich derer ein Algorithmus intransparent sein kann: Simulierbarkeit, Transparenz des Algorithmus und Zerlegbarkeit. Bei der Simulierbarkeit können Menschen jeden Schritt des Algorithmus nachvollziehen. Transparenz des Algorithmus bezieht sich bei Lipton auf die Funktionsweise und Eigenschaften des Algorithmus. Zerlegbarkeit bezieht sich auf die hier relevante Verständlichkeit. Sie erfordert, dass jeder Input und Output, jede Variable, jeder Parameter und jeder Rechenschritt intuitiv erklär- bzw. verstehbar ist. Dies ist bei Merkmalen insbesondere dann der Fall, wenn es alltagssprachliche Beschreibungen von ihnen gibt. Dieser Aspekt ist für die Frage der Diskriminierung von besonderer Bedeutung. Während eine einfache lineare Regression gut verstanden werden kann, fehlt diese

9 Pedreshi, Ruggieri und Turini (2008) gehören mit zu den Ersten, die Verfahren zur Entdeckung von Diskriminierung im Zusammenhang mit maschinellem Lernen entwickeln. Sie weisen bereits auf Schwierigkeiten hin, die sich ergeben, wenn scheinbar neutrale Merkmale nicht nur mit dem gesuchten Hauptmerkmal korrelieren, sondern auch mit diskriminierungsrelevanten Merkmalen. Dann muss, wie etwa Kamiran, Žliobaitė und Calders (2013) argumentieren, auch geprüft werden, ob diese Merkmale eine von der Korrelation zum diskriminierungsrelevanten Merkmal unabhängige Erklärungskraft für das Hauptmerkmal haben und ob die Korrelation zum diskriminierungsrelevanten Merkmal kausal ist.

Verstehbarkeit etwa bei hochdimensionalen, komplexen, wechselwirkenden und nichtlinearen Zusammenhängen oder neuronalen Netzwerken. Auch die Mustererkennung in Sensordaten (z. B. Bild- oder Spracherkennung) erzeugt komplexe Merkmale, die für Menschen oft keinen Sinn ergeben. Liest man etwa die Zwischenschichten neuronaler Netzwerke (die aus vielen Schichten zwischen Input und Output bestehen) aus, um zu sehen, wie die sensorischen Daten verarbeitet wurden und anhand welcher „Bilder" dann zum Beispiel Zahlen erkannt werden, so sieht man als Mensch nur Pixelhaufen, die nichts mit Zahlen zu tun zu haben scheinen (Burrell 2016).

Auch unverständliche Merkmale lassen sich, wenn auch nicht verlustfrei, in verständliche Merkmale überführen. Hierfür wurden gerade in den letzten Jahren verschiedene Verfahren entwickelt. Diese zielen zumeist darauf ab, die einflussstärksten Merkmale für einen bestimmten Output zu identifizieren. Hierbei können dieselben Verfahren, die eingesetzt werden, um unbekannte Merkmale zu identifizieren, auch eingesetzt werden, um unbekannte und unverständliche Merkmale in verständliche und bekannte Merkmale zu überführen.[10] Hierbei ist zu bedenken, dass diese Verfahren „nur" recht zuverlässige Mutmaßungen über die tatsächlich zur Anwendung gekommenen Merkmale anstellen. Diese Mutmaßungen können jedoch zuverlässiger sein als Vermutungen über Intentionen, die dem Verhalten von Menschen zugrunde liegen, da sie systematischer entwickelt werden können, indem Merkmale variiert werden und das Verhalten des Algorithmus in einer hohen Fallzahl untersucht werden kann.

Besonders schwierig ist es, Merkmale, die aus Sensordaten gewonnen werden, verständlich zu machen. Die Verarbeitung von Bilddaten etwa durch einen Convolutional-Neural-Network-Algorithmus führt zu Bewertungen oder Einstufungen, die nicht mehr für Menschen verständlich sind. Daher wurden z. B. Verfahren entwickelt, um diejenigen Stellen eines Bildes zu markieren, die für die Entscheidung oder Bewertung des Algorithmus besonders wichtig gewesen sind. Doch auch diese Verfahren können

10 Molnar (2019) diskutiert verschiedene aktuelle modellunspezifische Interpretationsverfahren (Verfahren, die unabhängig von dem jeweils eingesetzten Algorithmus angewandt werden können): LIME („local interpretable model-agnostic explanations"), PDP („partial dependence plot"), ICE („individual conditional expectation"), ALE („accumulated local effects"), „feature interaction", „feature importance", „shapely values" und „global surrogate". Diese Verfahren lassen sich je nach Ursache des Verständlichkeitsproblems unterschiedlich erfolgreich anwenden.

bei anderen Sensordaten scheitern. Bei der Verarbeitung von Sprachdaten kann zwar versucht werden, Verständlichkeit dadurch zu gewährleisten, dass nur Merkmale extrahiert werden, die Phonemen entsprechen (Ferragne, Gendrot, Pellegrini und Thomas 2019), doch kann Sprache auch anhand völlig anderer Merkmale analysiert werden, die nicht so leicht verständlich gemacht werden können. So existieren Fälle, bei denen Jobbewerber_innen allein aufgrund gesprochener Sätze ausgewählt werden. Anhand welcher Merkmale der Aussprache der Algorithmus Bewertungen vornimmt, kann völlig unverständlich sein. Selbstsicherheit, Bildungsgrad, Motivation und Ähnliches mögen sich erkennen lassen und verständliche Merkmale der Aussprache bilden. Aber es können auch Merkmale zur Entscheidung herangezogen werden, die sich nicht alltagssprachlich übersetzen lassen. Dies ist vergleichbar mit einem wissenschaftlichen Forschungsprozess, bei dem relevante Einflussfaktoren entdeckt werden, auf die alltagssprachlich nicht Bezug genommen wird und für die daher auch keine Ausdrücke existieren. Solche Merkmale werden dann im wissenschaftlichen Kontext mit neuen Namen belegt. Der Unterschied zum Algorithmus besteht jedoch darin, dass im wissenschaftlichen Kontext völlig transparent ist, was für Merkmale dies sind und welchen kausalen Einfluss diese auf ein Geschehen haben. Um diese Merkmale für Laien verständlich zu machen, ist vielleicht ein gewisser Aufwand nötig, aber es ist durchaus machbar. Extrem komplexe Merkmale, die aus Sensordaten extrahiert werden, können grundsätzlich für Menschen unverständlich sein. Somit besteht die Möglichkeit, dass es zu Ungleichbehandlungen anhand von Merkmalen kommt, die wir nicht hinreichend verständlich machen können. Die Thematisierung und Politisierung von Diskriminierungen anhand dieser Merkmale ist schwierig.

Bei dem Versuch, unverständliche Merkmale verständlich zu machen, indem man für den Output oder bestimmte Outputfälle einflussreiche verständliche Merkmale erzeugt, geht es auch darum, den Output erklärbar zu machen. Erklärbarkeit spielt eine wichtige Rolle für die Rechtfertigung von Ungleichbehandlungen und wird daher in 4.3 eingehender besprochen.

4.3 Fehlende Erklärungen

Selbst wenn die Merkmale, aufgrund derer der Algorithmus eine Ungleichbehandlung vornimmt, bekannt sind und diese Merkmale verständlich sind, kann es sein, dass nicht beurteilt werden kann, ob eine Diskriminierung vorliegt. Dies liegt daran, dass Ungleichbehandlungen grundsätzlich auch angemessen sein können und wir Erklärungen für die Ungleichbehandlung benö-

tigen, um entscheiden zu können, ob die Ungleichbehandlung angemessen war oder nicht. Erklärungen ergeben sich nicht schon aus der Kenntnis der verwendeten Merkmale. Angemessen ist eine Ungleichbehandlung insbesondere dann, wenn die Prognosen zutreffend sind, die Stellvertretermerkmale kausal für das Hauptmerkmal sind und es normativ akzeptiert wird, dass diese Merkmale für den Entscheidungskontext verwendet werden. Erklärungen helfen uns dabei, Prognosen besser einzuschätzen zu können, zu beurteilen, ob tatsächlich ein Kausalzusammenhang besteht, und Merkmale zu identifizieren, die der Algorithmus nicht verwendet hat, die aber für den Zusammenhang von Stellvertretermerkmal und Hauptmerkmal wesentlich sind. Diese Aspekte sind nicht nur wichtig, um entscheiden zu können, ob Benachteiligungen sachlich angemessen sind, sondern auch, um beurteilen zu können, ob sie sachbezogen normativ angemessen sind.

Im Versicherungskontext können unter bestimmten Umständen etwa dann Ungleichbehandlungen nicht diskriminierend sein, wenn diese gute Prognosen liefern, da sie auf der Grundlage einer „relevanten und [auf] genauen versicherungsmathematischen und statistischen Daten beruhenden Risikobewertung“ stattfinden (Britz 2008, 161, und Avraham 2018). Die Erklärung für die Ungleichbehandlung besteht damit im Verweis auf die historisch erhobenen Daten über Versicherungsfälle und die statistischen Methoden zur Datenauswertung. Weiterhin verfügen wir oft über intuitive Erklärungen darüber, weshalb ein Merkmal der Ungleichbehandlung für Entscheidungen kausal relevant ist. Dass etwa Arbeitslosigkeit und Armut ein Kreditausfallrisiko zur Folge haben, ist wenig überraschend. Kennen wir diese Zusammenhänge und Erklärungen, können wir deutlich besser normative Beurteilungen vornehmen oder werden überhaupt erst in die Lage versetzt, normative Beurteilungen vorzunehmen. Wenn es etwa ärmeren Menschen erschwert wird, aus der Armut herauszukommen, wenn sie keine Kredite aufnehmen können, könnten wir es als normativ wichtig erachten, diese Menschen vor zu starken Benachteiligungen bei der Kreditvergabe zu schützen und diese Benachteiligungen daher als diskriminierend auffassen.

Während also in klassischen versicherungsmathematischen Verfahren der Zusammenhang zwischen historischen Daten und dem Modell, das für Ungleichbehandlungen verwendet wird, verhältnismäßig transparent und intuitiv plausibel ist[11] und ein Streit um die sachliche und sachbezogen

11 Avrahams (2018) Artikel zeigt, dass in klassischen Versicherungsfällen statistische Daten, Merkmale und Kausalzusammenhänge recht transparent sind,

normative Angemessenheit relativ klar geführt werden kann, stellt sich dies im Fall intransparenten maschinellen Lernens anders dar. Weder verfügen wir über intuitive Erklärungen für den Zusammenhang zwischen Stellvertretermerkmal (soweit wir dieses Merkmal überhaupt kennen) und Hauptmerkmal, die als Startpunkt für weitere Untersuchungen dienen könnten, noch erkennen wir den Zusammenhang zwischen historischen Trainingsdaten und der Verwendung und Gewichtung der Stellvertretermerkmale. Ohne diese Erklärungen ist es jedoch nicht erkennbar, ob die Ungleichbehandlung sachlich angemessen ist, und auch normative Überlegungen werden ohne diese Erklärungen deutlich erschwert. Es bleibt unklar, ob ein relevanter Kausalzusammenhang zwischen Stellvertretermerkmal und Hauptmerkmal besteht, wie die Gewichtung und Interaktion der Merkmale begründet ist und ob die Trainingsdaten und die Lernmethode geeignet waren, um das Erlernen von diskriminierenden Stellvertretermerkmalen zu vermeiden. Wenn die Trainingsdaten etwa historische Diskriminierungen enthalten oder in Bezug auf entscheidende Merkmale nicht ausgewogen und repräsentativ waren, dann kann es leicht passieren, dass Korrelationen erlernt werden, die der Sache nicht gerecht werden und zu diskriminierenden Ungleichbehandlungen führen.

Der Verweis auf genaue statistische Methoden oder eine gute Prognoseleistung eines Algorithmus reicht also nicht aus. Dies liegt nicht nur daran, dass eine gute Prognose noch immer falsche Bewertungen enthält und damit Einzelfallgerechtigkeit nicht herstellen kann.[12] Selbst eine korrekte Prognose kann problematisch und diskriminierend sein, etwa wenn eine Scheinkorrelation für die gute Prognose verantwortlich ist, aber kein kausaler Zusammenhang besteht. Während etwa eine Scheinkorrelation zwischen genetischen Merkmalen und der Entwicklung von Krankheiten eine Ungleichbehandlung in Bezug auf Krankenversicherungen nicht rechtfertigen dürfte und diskriminierend sein könnte, ist dies bei tatsächlichen kausalen Zusammenhängen zumindest umstritten (z. B. Lippert-Rasmussen 2014,

aber ein komplizierter Streit darum geführt wird, inwiefern die Verwendung der auf statistischen Daten beruhenden Merkmale sachlich und sachbezogen normativ angemessen ist. Die Möglichkeit, diesen Streit zu führen, muss im Fall intransparenter Diskriminierung durch maschinelles Lernen erst erzeugt werden.

12 Ein besonderes Problem stellt hierbei auch die Abwägung zwischen Datenschutz und Einzelfallgerechtigkeit dar. Die Prognoseleistung wird schließlich umso genauer, je mehr Daten über die betroffenen Personen bekannt sind.

95ff.). Nun könnte man einwenden, dass dies nicht relevant sei, solange die Prognose korrekt war. Damit scheint es vielleicht so, als läge eine „all things considered" unproblematische Diskriminierung vor, die keine negativen Folgen hat, weil die Ungleichbehandlung zwar aufgrund falscher Annahmen stattfindet, aber im Endeffekt zum richtigen Ergebnis geführt hat. Das Vertrauen in das Modell und die Anwendbarkeit auf zukünftige Sachverhalte dürften damit aber erschüttert sein. Wenn der Algorithmus keine Kausalbeziehung verwendet, sondern nur eine Scheinkorrelation, kann bei einem Wechsel des Anwendungskontexts plötzlich eine extrem hohe Fehlerquote entstehen. Wird dies als ein ernsthaftes Problem gesehen, ist die Diskriminierung „all things considered" nicht unproblematisch oder gerechtfertigt.

Selbst wenn wir davon ausgehen, dass die Korrelation aufgrund eines Kausalzusammenhangs besteht, bleiben einige Fragen offen.

- Wie sind die Merkmale gewichtet? So könnten wir es akzeptieren, dass die Nichtbedienung vergangener Schulden für Kreditscoring herangezogen wird. Aber eine so starke Gewichtung selbst kleiner nicht bezahlter Schulden, dass etwa das Abschließen von Handyverträgen oder die Eröffnung eines Kontos unmöglich wird, scheint sachlich nicht angemessen.
- Über welchen Mechanismus sind Stellvertreter- und Hauptmerkmal verbunden? Hier könnte sich herausstellen, dass der Mechanismus Merkmale beinhaltet, die im Datensatz nicht repräsentiert werden, deren Einfluss in diesem Kontext jedoch diskriminierend ist. Dann läge ein Fall indirekter Diskriminierung vor, den wir leicht erkennen können, wenn wir den Mechanismus zwischen Stellvertreter- und Hauptmerkmal kennen. Wenn sich etwa herausstellt, dass der Bildungsabschluss in einem Unternehmen nicht deshalb mit der Jobperformance korreliert, weil größere Sachkompetenz vorliegt, sondern weil Arbeitnehmer_innen mit niedrigem Bildungsabschluss gemobbt werden, dann ist eine Bewerber_innenauswahl auf der Grundlage von Bildungsabschluss diskriminierend.
- Warum interagieren bestimmte Merkmale miteinander? Auch hier kann uns die Antwort auf Mechanismen hinweisen, die diskriminierungsrelevante Merkmale beinhalten.
- Wodurch erklärt sich ein nichtlinearer Zusammenhang der Merkmale? Es könnte sein, dass Merkmal M1 eigentlich linear mit dem Hauptmerkmal zusammenhängt, aber unter Anwesenheit von M2 ab einem gewissen Schwellenwert plötzlich negativ mit dem Hauptmerkmal

korreliert. Auch hier ist die Kenntnis des Mechanismus relevant für die Identifikation potentieller indirekter Diskriminierungen.

- Wird das Vorliegen des Stellvertretermerkmals durch ein Merkmal verursacht, dessen Verwendung sachlich unangemessen wäre? Die Bereitschaft, Überstunden zu absolvieren, könnte zwar für Jobperformance kausal relevant sein, aber durch nach Geschlecht ungleich verteile Hausarbeit und Kinderbetreuung erklärbar sein. Auch hier läge eine indirekte Diskriminierung vor.

Dies sind Fragen, die sich auch bei herkömmlichen statistischen Verfahren stellen. Je intransparenter und komplexer die verwendeten Merkmale und statistische Verfahren jedoch werden, desto schwieriger wird es, diese Fragen zu beantworten.[13] Hinweise auf Antworten finden sich oft mit Blick auf die Trainingsdaten und die Untersuchung, wie diese erhoben wurden, ob sie repräsentativ sind oder vorbelastet. Verschiedene Verfahren des maschinellen Lernens können auch indirekte Diskriminierungen in Datensätzen entdecken und vermeiden. Dies setzt jedoch voraus, dass alle Merkmale bekannt sind, die auf ihren diskriminatorischen Charakter hin untersucht werden sollen. Dies schließt insbesondere auch diejenigen Merkmale ein, die nicht im Datensatz repräsentiert sind und die im Hinblick auf indirekte Diskriminierung überprüft werden sollen. Ist dies nicht der Fall, sind wir für unsere menschlichen Überlegungen auf Erklärungen angewiesen, die komplexe Algorithmen des maschinellen Lernens und selbst XAI nicht immer liefern können.

5. Fazit

Ungleichbehandlungen auf der Grundlage maschinellen Lernens haben das Problem, dass nicht unmittelbar transparent ist, anhand welcher Merkmale warum eine Ungleichbehandlung stattfindet. Bisherige Ansätze beschränken sich darauf, verständliche Merkmale zu identifizieren oder zu erzeugen, die unter explizitem Diskriminierungsschutz stehen. Damit geraten jedoch neue Diskriminierungen aus dem Blick, die in Zukunft erhebliche Auswirkungen

13 Ein großes Problem stellt auch die aus Kostengründen stattfindende Verwendung von vortrainierten Algorithmen dar. So kann ein Algorithmus etwa zur Bilderkennung trainiert werden und dann für andere Anwendungskontexte genutzt werden. „Transferable AI“ können zwar gute Prognose- oder Klassifikationsleistungen erbringen, sind jedoch im Hinblick auf Erklärbarkeit und verwendete Trainingsdaten hochgradig intransparent.

haben können. Ich habe gezeigt, dass es plausibel ist, sachlich unangemessene als auch sachbezogen normativ unangemessene Benachteiligungen anhand intransparenter (neuer) Merkmale und Verfahren als Diskriminierung zu verstehen. Weiterhin hat sich gezeigt, dass unterschiedliche Aspekte dieser Intransparenz unterschiedliche Konsequenzen haben und spezifische Lösungsstrategien zur Identifikation und Beseitigung von Diskriminierung erfordern. Unbekannte Merkmale müssen identifiziert werden können, unverständliche Merkmale hinreichend verständlich gemacht werden können und Korrelationen und Trainingsverfahren erklärbar sein. Anschließend müssen die diskriminierenden Elemente des Algorithmus beseitigt werden können. Insofern keine angemessenen Lösungsstrategien vorliegen, muss darüber nachgedacht werden, für bestimmte Kontexte nur Formen des maschinellen Lernens zuzulassen, die sich hinreichend auf diskriminierende Konsequenzen untersuchen lassen.

Literatur

Altman, Andrew. 2011. „Discrimination“. In *The Stanford Encyclopedia of Philosophy,* herausgegeben von Edward N. Zalta, URL = https://plato.stanford.edu/archives/spr2011/entries/discrimination/.

Altman, Andrew. 2016. „Discrimination“. In *The Stanford Encyclopedia of Philosophy,* herausgegeben von Edward N. Zalta, URL = https://plato.stanford.edu/archives/win2016/entries/discrimination/.

Avraham, Ronen. 2018. „Discrimination and insurance“. In *The Routledge Handbook of the Ethics of Discrimination,* herausgegeben von Kasper Lippert-Rasmussen, 335–347.

Britz, Gabriele. 2008. *Einzelfallgerechtigkeit versus Generalisierung,* Tübingen: Mohr Siebeck.

Buolamwini, J., und T. Gebru. 2018. „Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification“, *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, in PMLR 81:77–91.

Burrell, J. 2016. „How the machine ‘thinks’: Understanding opacity in machine learning algorithms“. *Big Data & Society,* 3(1), 1–12.

Custers, B., T. Calders, B. Schermer und T. Zarsky (Hrsg.). 2013. *Discrimination and Privacy in the Information Society*. Berlin Heidelberg: Springer.

Dastin, J. 2018. *Amazon scraps secret AI recruiting tool that showed bias against women, Reuters,* abgerufen von: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G.

Eidelson, Benjamin. 2015. *Discrimination and Disrespect,* Oxford: Oxford University Press.

Ferragne, Emmanuel, Cédric Gendrot und Thomas Pellegrini. 2019. „Towards phonetic interpretability in deep learning applied to voice comparison". In *Proceedings of the 19th International Congress of Phonetic Sciences, Melbourne, Australia 2019,* herausgegeben von Sasha Calhoun, Paola Escudero, Marija Tabain und Paul Warren, 790–794. Canberra, Australia: Australasian Speech Science and Technology Association Inc.

FRA (European Union Agency for Fundamental Rights). 2018. *#BigData: Discrimination in data-supported decision making*. https://fra.europa.eu/sites/default/files/fra_uploads/fra-2018-focus-big-data_en.pdf, abgerufen am 06.03.2020.

Fuster, A., P. Goldsmith-Pinkham, T. Ramadorai und A. Walther. 2018. *Predictably Unequal? The Effects of Machine Learning on Credit Market*. SSRN: https://ssrn.com/abstract=3072038 oder http://dx.doi.org/10.2139/ssrn.3072038, abgerufen am 06.03.2020.

Garcia, J. L. A. 2018. „Discrimination and Virtue". In *The Routledge Handbook of the Ethics of Discrimination,* herausgegeben von Kasper Lippert-Rasmussen, 174–182.

Kamiran, F., I. Žliobaitė und T. Calders. 2013. „Quantifying explainable discrimination and removing illegal discrimination in automated decision making". *Knowledge and information systems,* 35(3), 613–644.

Khaitan, T. 2015. *A theory of discrimination law,* Oxford: Oxford University Press.

Lippert-Rasmussen, Kasper. 2014. *Born Free and Equal?* Oxford: Oxford University Press.

Lippert-Rasmussen, Kasper 2016. „Discrimination, Freedom, and Intentions". *The Modern Law Review,* 79(5), 901–918.

Lippert-Rasmussen, Kasper (Hrsg.). 2018. *The Routledge Handbook of the Ethics of Discrimination,* London and New York: Routledge.

Lipton, Z. C. 2017. *The mythos of model interpretability*. arXiv preprint arXiv:1606.03490.

Molnar, Christoph. 2019. *Interpretable Machine Learning. A Guide for Making Black Box Models Explainable,* https://christophm.github.io/interpretable-ml-book/.

Mozur, P. 2019. *One Month, 500,000 Face Scans: How China Is Using A.I. to Profile a Minority,* The New York Times, abgerufen von: https://www.nytimes.com/2019/04/14/technology/china-surveillance-artificial-intelligence-racial-profiling.html.

Pedreshi, D., S. Ruggieri und F. Turini. 2008. „Discrimination-aware data mining". In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining,* 560–568. ACM.

Thomsen, Frej Klem. 2013. „But Some Groups Are More Equal Than Others: A Critical Review of the Group-Criterion in the Concept of Discrimination". *Social Theory and Practice,* Vol. 39, No. 1, 120–146.

Stoljar, Natalie. 2018. „Discrimination and Intersectionality". In *The Routledge Handbook of the Ethics of Discrimination,* herausgegeben von Kasper Lippert-Rasmussen, 68–79.

David Wasserman (1998): „Discrimination, Concept of". In *Encyclopedia of Ethics,* herausgegeben von R. Chadwick, 805–814. San Diego, CA: Academic Press.