Hochschule Düsseldorf
University of Applied Sciences

Fachbereich Medien
Faculty of Media

Datenbanken und E-Business
Databases and E-Business

Prof. Dr.-Ing. Thomas C. Rakow
thomas.rakow@hs-duesseldorf.de
dbe.medien.hs-duesseldorf.de

Diskriminierung im Machine Learning und Erklärbarkeit von Algorithmen

von

Arabella Jackszis

Arbeitspapier des Lehrgebiets
Datenbanken und E-Business
No. 2/ 2021

herausgegeben von

Thomas C. Rakow

Bild: "Hand mit Zeichenstift" von Klaus Kammerichs, Münsterstraße 156, Düsseldorf, Thomas Rakow, September 2019.

Zum Geleit

In E-Shops werden aufgrund vielfältig gesammelter Nutzungsdaten Kaufempfehlungen gegeben. Soziale Netzwerke stellen neue Kontakte mittels der Nutzereigenschaften her. Lernende Systeme werden zur Markierung von Fake-News verwendet. Bei der Anwendung KI-basierter Systeme für diese Aufgaben wurde beobachtet, dass es zu einer unerwünschten Gender- oder Rassendiskriminierung kommen kann. Problematisch ist auch die Anwendung, wenn im Falle einer Produkthaftung der Nachweis der Korrektheit nicht beigebracht werden kann.

In dieser Arbeit werden aufgrund einer Literaturrecherche die Methoden zur erklärenden Anwendung von Algorithmen analysiert. Auf welcher Grundlage basieren die Algorithmen und sind diese transparent für den Nutzer? Inwieweit werden rechtliche und ethische Standards eingehalten? Insbesondere die Anwendung der weit verbreiteten neuronalen Netze zum Machine Learning wird dargestellt und eingeordnet in die Fragestellung.

Frau Jackszis hat die Literatur zu diesem Thema in vorbildlicher Weise recherchiert. Sie ordnet die Begrifflichkeiten und entwickelt eine Darstellung der Risiken zur Diskriminierung in den einzelnen Schritten des Machine Learning. Sie stellt die gesetzlichen und ethischen internationalen Standards gegen Diskriminierung sowie Erklärbarkeit übersichtlich dar.

Als praktischen Anteil ihrer Arbeit hat Frau Jackszis eine Dashboard-Anwendung entwickelt, um die Wirkungsweisen von Diskriminierung in Verfahren des Machine Learning interaktiv darzustellen. Diese Arbeit soll an der Hochschule in der Lehre eingesetzt werden.

Ich danke Frau Jackszis, dass sie diesen Teil ihrer Masterarbeit[1] in den Arbeitspapieren des Lehrgebiets Datenbanken und E-Business publiziert und ihre Entwicklung der Hochschule zur Verfügung gestellt hat.

Düsseldorf, 29.01.2021 *Professor Dr.-Ing. Thomas C. Rakow*

[1] Jackszis, A.: Diskriminierung im Machine Learning und Erklärbarkeit von Algorithmen. Masterthesis, FB Medien, HS Düsseldorf, Dezember 2020. Erstprüfer: Thomas C. Rakow, Zweitprüfer: Björn Salgert.

Inhaltsverzeichnis

Kurzfassung

Künstliche Intelligenz und Machine Learning sind hochaktuelle Themen in der Informatik. Fortschrittliche Technologien und Vorgehensweisen schaffen neue Einsatzfelder für Anwendungen, die Verhaltensweisen des Menschen erlernen und nachahmen. Dieses innovative Vorgehen birgt jedoch einige Gefahren. Das menschliche Verhalten ist nicht immer objektiv und korrekt. Oft beherrschen Vorurteile das Treffen menschlicher Entscheidungen, sodass Personen diskriminiert werden. Es besteht das Risiko, dass Maschinen diese Verhaltensweise übernehmen.

Hinzu kommt die fehlende Erklärbarkeit der Algorithmen. Je nach Modell ist nicht eindeutig, nach welchen Prinzipien eine Entscheidung getroffen wird. Das erschwert das Aufdecken von Diskriminierung und senkt die Vertrauenswürdigkeit des Systems.

In dieser Arbeit werden Ursachen für Diskriminierung und fehlende Transparenz, sowie Lösungsmethoden erarbeitet. Dabei wird erörtert, inwieweit die beiden Themen aufeinander einspielen und gemeinsam zur Qualität der Software beitragen.

Abstract

Artificial Intelligence and Machine Learning are matters of high interest in Information Technology. Advanced technologies and approaches are creating new areas of use for applications that study and imitate human behavior. At the same time, this innovative approach poses new dangers. Human behavior is not always objective and correct. Prejudice often dominates human decisions and therefore leads to discrimination. There is a risk that machines adopt this behavior.

On top of this, algorithms are often not explainable. Depending on the model, the principles that are relevant for the decision are not clear. This

makes it harder to detect discrimination and reduces the trustworthiness of a system.

In this paper, causes of bias in Machine Learning and the lack of transparency, as well as methods to eliminate these problems, are elaborated. It works out the interaction between the two issues and how they contribute to the quality of the software.

Stichworte

Diskriminierung, Fairness, Erklärbarkeit, Transparenz, Nachvollziehbarkeit, Machine Learning, künstliche Intelligenz

Anzahl Wörter

14.332

1. Einleitung

Ein Auto entscheidet, ob es ein Kind oder einen Rentner überfährt (Simmank, 2018). Analysetools unterstützen Lehrende im Benotungsprozess, geben dabei aber schlechtere Empfehlungen für Schüler mit Migrationshintergrund (Wehner & Köchling, 2020). Ein AI-System des Unternehmens Amazon analysiert Bewerbungen und stuft weibliche Bewerber als ungeeignet ein (Vincent, 2018). Künstliche Intelligenz hat ein breites Einsatzgebiet und schafft Möglichkeiten zur Automatisierung aufwändiger Prozesse. Die Beispiele zeigen jedoch die große Verantwortung der Modelle, denn die Folgen mangelnder Qualität sind schwerwiegend.

Vertrauenswürdigkeit von maschinellen Entscheidungen gewinnt seit einigen Jahren zunehmend an Bedeutung. Die Arbeit erörtert Ursachen für diskriminierende Entscheidungen und stellt Methoden zur Prävention und Nachbearbeitung vor. Es wird analysiert, inwiefern Transparenz zur Vermeidung und Aufdeckung von Diskriminierung beiträgt. In der Fachliteratur werden Fairness und Erklärbarkeit oft getrennt voneinander betrachtet (Sharma, et al., 2019). Diese Arbeit verfolgt das Ziel, den Zusammenhang und die resultierenden Vorteile zu erörtern.

1.1. Motivation

1959 wurden erste Systeme entwickelt, die nicht explizit programmiert waren, sondern einen Lernprozess durchliefen (Munoz, 2014). Seitdem hat sich maschinelles Lernen enorm entwickelt, insbesondere durch die Möglichkeit, hohe Datenmengen zu verarbeiten und analysieren. Anhand dessen werden präzise Prognosen und Klassifizierungen vorgenommen.

Diese Entwicklungen schaffen großes Potenzial für die Automatisierung und Präzisierung vieler Anwendungsgebiete. In der Medizin, beim autonomen Fahren, im Recht und in vielen weiteren Anwendungsgebieten gibt

es dadurch die Möglichkeit, den Menschen bei Entscheidungen zu unterstützen und sogar neue wissenschaftliche Erkenntnisse zu erarbeiten.

Maschinelle Entscheidungen können sich allerdings auch negativ auf die Gesellschaft auswirken, wenn sie unüberlegt und unüberwacht eingesetzt werden. Nach wie vor müssen grundrechtlich geschützte Werte wie die Menschenwürde geachtet werden (Orwat, 2019). Wenn Algorithmen geschützte Personengruppen systematisch benachteiligen, verstößt das nicht nur gegen das Grundgesetz, sondern auch gegen jegliche ethische Leitlinien.

Die Auswirkungen von Diskriminierung sind an dem System COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) zu erkennen (Orwat, 2019). Es dient der Bewertung von Straftätern, um zu bewerten, ob sie potenziell in der Zukunft erneut straffällig werden. Diese Bewertung wird anhand persönlicher Merkmale vorgenommen, um Richterinnen und Richter in Teilen der USA in dem Urteil zu unterstützen, ob die angeklagte Person vorzeitig entlassen werden kann. Untersuchungen zufolge liefert das System schlechtere Ergebnisse für Afro-Amerikanische Angeklagte. Da das Modell für die Öffentlichkeit nicht transparent ist, kann weder die Ursache, noch der Grad der Diskriminierung sicher festgestellt werden (Angelino, et al., 2017).

Das zeigt die Auswirkungen von KI-Systemen auf die Gesellschaft und Individuen. Gleichzeitig steigt durch fehlende Transparenz das Misstrauen in automatisierte Prozesse und Entscheidungen von Unternehmen und Regierungen. Deshalb muss zur Qualitätssicherung im Entwicklungsprozess besonders auf Fairness und Erklärbarkeit geachtet werden. Dadurch können vertrauenswürdige Modelle entwickelt werden, die das Potenzial des Machine Learnings voll ausschöpfen.

1.2. Überblick über das Projekt

Zunächst werden die wissenschaftlichen Grundlagen über Diskriminierung und Erklärbarkeit erarbeitet. Anschließend wird der Zusammenhang der beiden Themenbereiche dargestellt.

Die Ausführungen über Diskriminierung zeigen, dass viele verschiedene Faktoren auf das Modell einwirken. Entwickler, ökonomische, und organisatorische Umstände formen das Ergebnis und die Auswirkungen auf betroffenen Personen. Gleichzeitig existieren einige Maßnahmen, Diskriminierung präventiv zu unterbinden oder im Nachhinein zu eliminieren.

Auch für die Förderung der Transparenz ermöglichen einige Methoden einen Einblick in die Entscheidungsfindung der Algorithmen. Da sich durch diese beiden Themen bedeutende Vorteile für Entwickler und Anwendende ergeben, wird in dieser Arbeit deren Zusammenhang hervorgehoben.

Kapitel 2 beinhaltet die Ausführungen über Diskriminierung. Anschließend folgen die Erkenntnisse zu Erklärbarkeit der Algorithmen und in Kapitel 4 deren Zusammenhang. Im Anschluss werden Schlussfolgerungen für die Softwareentwicklung im Machine Learning gezogen.

2. Fairness

Beim Einsatz einer neuen Software sind die zwei Qualitätsmerkmale Fairness und Erklärbarkeit von großer Bedeutung. Modelle, die fair und erklärbar sind, verhindern negative Auswirkungen auf die Gesellschaft. Die in Kapitel 1 genannten Beispiele zeigen, dass Machine Learning aktiv in verschiedenen Lebensbereichen eingesetzt wird und das Leben Vieler beeinflusst. Aufgrund von fehlender Transparenz über die Entscheidungskriterien des Modells ist es zumeist nicht möglich, Objektivität zu garantieren und die Folgen maschinell getroffener Entscheidungen abzuschätzen.

Die hier vorgestellten Erkenntnisse wurden durch eine umfangreiche Literaturrecherche gewonnen. Über die wissenschaftliche Suchmaschine *Google Scholar* (www.scholar.google.de) können Fachliteratur und aktuelle Veröffentlichungen anhand von Schlagwörtern und Filter- oder Sortieroptionen lokalisiert werden. Für die Suche wurden folgende Begriffe verwendet: „bias machine learning“, „fairness artificial intelligence“, „neural networks fairness“, „explainable AI“ und „machine learning explainability“.

Für die Recherche wurden möglichst aktuelle Veröffentlichungen zu Rate gezogen, da das hier betrachtete Thema erst seit wenigen Jahren an Bedeutung zunimmt und immer wieder neue Erkenntnisse vorgestellt werden. Darüber hinaus entwickeln sich die Technologien und Verfahrensweisen für künstliche Intelligenz sehr schnell.

In Kapitel 2.1 und 2.2 erfolgt die Definition der relevanten Begriffe im Machine Learning. Um nachvollziehen zu können, welche Forderungen an KI-Systeme gestellt werden, wird in Kapitel 3 der rechtliche Rahmen abgesteckt. Kapitel 2.4 und 2.5 verfolgen das Ziel, beurteilen zu können, welche Faktoren zur Entstehung von Diskriminierung beitragen und welche Methoden zur Verbesserung der Fairness beitragen.

2.1. Allgemeines über Machine Learning

Maschinelles Lernen beschreibt die Entwicklung eines Algorithmus, der aus vorgegebenen Daten ein Muster erkennt, um ein bestimmtes Verhalten zu übernehmen und selbst anzuwenden (Frochte, 2019, p. 13). Anhand dieser Daten „erlernt“ der Algorithmus die Entscheidungsgrundlagen aus der darin repräsentierten Realität und ist im Anschluss selbst in der Lage, neue Datenobjekte zu bewerten, das heißt eine Prognose zu machen. Eine Prognose ist beispielsweise die Schätzung eines Hauspreises. Die Daten für das Training des Modells bestehen in diesem Fall aus einer Menge von Häusern mit verschiedenen Eigenschaften, die ausschlaggebend für den Preis des Hauses sind (beispielsweise „Ort“, „Anzahl der Räume“, oder „Quadratmeter“). Diese ermöglichen es dem Algorithmus, zu erkennen, von welchen Faktoren der Preis abhängig ist.

Im Machine Learning wird zwischen überwachtem und unüberwachtem Lernen unterschieden (Kolodiazhnyi, 2020). Beim überwachten Lernen erfolgt das Training des Modells anhand von möglichst umfangreichen Daten bestehend aus Attributen (hier „Features“ genannt) und entsprechenden Werten. Dazu gehören zum Beispiel Angaben wie Alter, Schulabschluss oder Herkunft. Jedem Datum ist außerdem die korrekte Ausgabe zugeordnet (hier Label genannt). Im oben genannten Beispiel ist das Label der tatsächliche Preis des Hauses. Anhand dieser sogenannten Trainingsdaten wird das Modell trainiert. Der Algorithmus minimiert anhand des vorgegebenen Labels die Abweichung zur eigenen Prognose.

In den folgenden Kapiteln wird in den meisten Fällen auf künstliche neuronale Netze für eine Klassifizierung Bezug genommen (Abbildung 1) Das gesamte Netz ist in mehrere Schichten (Layer) aufgeteilt, die jeweils eine Menge an Knotenpunkten enthalten. Diese werden „Neuronen“ genannt, da sich die Funktionsweise am menschlichen Gehirn orientiert, das auf ähnliche Weise aus Fehlern lernt und sich anpasst, um diese in Zukunft zu vermeiden (Kolodiazhnyi, 2020, p. 313). Die Schichten sind unterteilt in

Input Layer, mehrere Hidden Layers und ein Output Layer. Im Folgenden werden diese Schichten genauer erläutert.

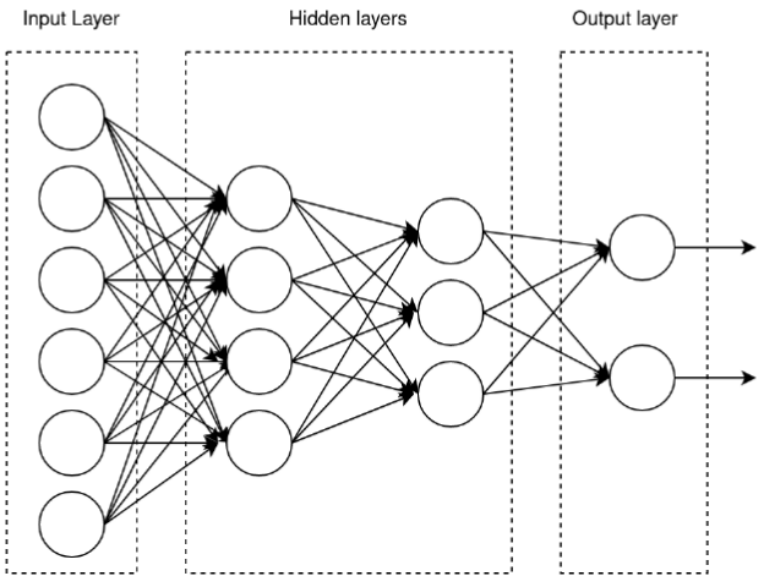

Abbildung 1: Aufbau eines neuronalen Netzes.Quelle: (Kolodiazhnyi, 2020, p. 320)

Die Neuronen aus dem Input Layer entsprechen den Features aus den Trainingsdaten. Durch die Berechnung der gewichteten Summe aller Neuronen aus dieser Schicht werden die Neuronen in den darauffolgenden Schichten aktiviert (Abbildung 2). Neuronen (x_1,x_2, x_i und x_n), die als Input für das betrachtete Neuron dienen, werden jeweils mit einem Gewicht w belegt und anschließend aufsummiert. Durch den zusätzlichen Summanden (w_0) kann die Summe verschoben werden. Daraufhin wird eine Aktivierungsfunktion $f(sum)$ auf das Neuron angewandt, um die weitere Verarbeitung – beispielsweise durch eine Normalisierung – zu ermöglichen.

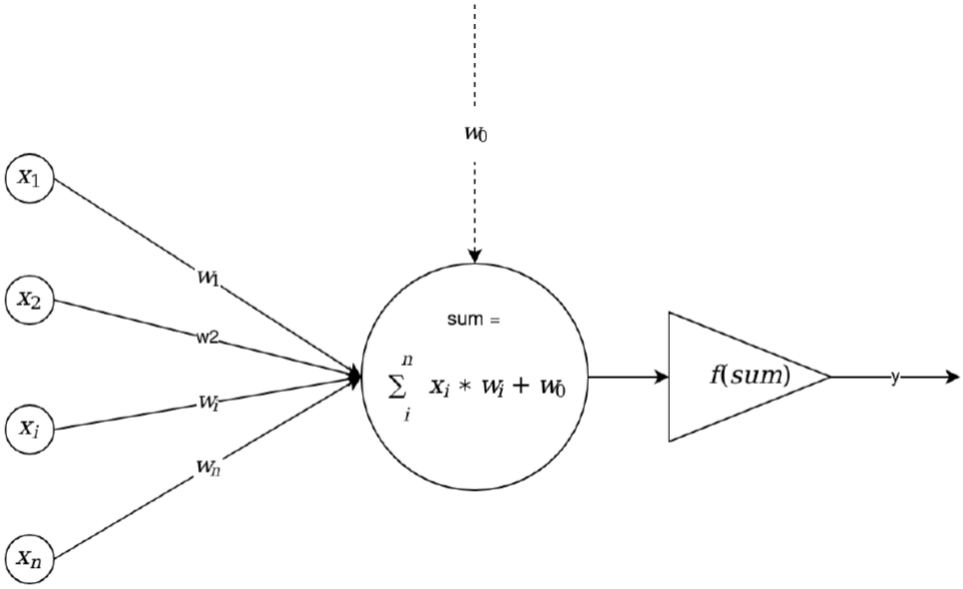

Abbildung 2: Neuron mit Input, Verarbeitung und Output. Quelle: (Kolodiazhnyi, 2020, p. 315)

Das Output Layer entspricht der Prognose, die das Modell treffen soll. Die Belegung der Neuronen in dieser Schicht gibt die Wahrscheinlichkeit an, dass die entsprechende Kategorie die korrekte Prognose ist.

Während des Trainings des neuronalen Netzes stehen besonders die Gewichtungen w der Features im Fokus. Diese müssen so gewählt werden, dass das Ergebnis möglichst nah am Label im Trainingsdatensatz ist. Für diese Optimierung gibt es verschiedene Algorithmen, beispielsweise den „Gradient Descent“ Algorithmus. Dabei geht es um die Minimierung der Abweichung der vom Modell berechneten Prognose vom Label. Die Gewichtungen werden immer wieder angepasst, um so schrittweise ein genaues und treffsicheres Modell zu entwickeln (Kolodiazhnyi, 2020, p. 17). Dieses ist anschließend in der Lage, Prognosen für unbekannte Datenobjekte zu treffen.

2.2. Definition des Begriffs „Fairness“

In der Europäischen Union und der Bundesrepublik Deutschland gilt der Begriff Diskriminierung als „benachteiligende, ungerechtfertigte Ungleichbehandlung von Personen in Anknüpfung an ein geschütztes Merkmal“ (Orwat, 2019, p. 24). Welche Merkmale gesetzlich als geschützt gelten, wird in Kapitel 2.3 vorgestellt.

Häufig tritt in dem hier betrachteten Zusammenhang der englische Begriff „Bias“ auf, der aus dem Englischen heraus in „Vorurteil“, „Neigung“ oder „beeinflusst sein“ übersetzt werden kann (Cambridge University Press). Im technischen Kontext benachteiligt oder bevorzugt eine Software bestimmte Personengruppen demnach systematisch. Darunter können computerbasierte Entscheidungen, Prognosen oder andere Zuweisungen fallen (Friedman & Nissenbaum, 1996).

Teilweise wird der Begriff jedoch etwas weiter gefasst. So kann „Bias“ auch den Umstand beschreiben, dass die Prognosen eines Modells systematisch vom korrekten Ergebnis abweichen (Baer, 2019). In dieser Arbeit wird jedoch vorrangig die erstere Definition verwendet: Die Ungleichbehandlung von Personen auf Basis eines geschützten Merkmals.

2.3. Rechtliche und ethische Standards

Der Begriff „Künstliche Intelligenz“ ist schon seit den 1950er Jahren präsent. Bis heute gibt es jedoch keine einheitliche, klare Definition des Begriffs „Intelligenz“ in diesem Bereich (Zielke, 2020). Neue Anwendungsszenarien verursachen neue Herausforderungen, die sich negativ auswirken. Es reicht nicht, bestehende Standards auf das neue Szenario zu übertragen –angepasste Richtlinien und Gesetze sind erforderlich. Trotzdem müssen bestehende Gesetze nach wie vor in allen Phasen der Softwareentwicklung berücksichtigt werden. Im Folgenden werden Gesetze

und Standards aus Deutschland, der Europäischen Union und aus den Vereinigten Staaten von Amerika beleuchtet (Abbildung 4).

Im Grundgesetz ist verankert, dass „Niemand […] wegen seines Geschlechts, seiner Abstammung, seiner Rasse, seiner Sprache, seiner Heimat und Herkunft, seines Glaubens, seiner religiösen oder politischen Anschauungen benachteiligt oder bevorzugt werden [darf]. Niemand darf wegen seiner Behinderung benachteiligt werden.“ (Art. 3 Abs. 3 GG).

Seit dem 18.08.2006 gilt darüber hinaus das Allgemeine Gleichbehandlungsgesetz, in dem die zu schützenden Merkmale um Alter und sexuelle Identität erweitert werden. Ziel ist die Verhinderung und Beseitigung von Benachteiligungen (§1 AGG). Das Gesetz findet insbesondere beim Zugang zu Arbeit und Bildung Anwendung, sowie im Sozialschutz, bei sozialen Vergünstigungen, Dienstleistungen und Gütern, wie beispielsweise dem Wohnraum (§2 Abs. 1 AGG). Da in vielen dieser Bereiche Machine Learning zum Einsatz kommt, ist das Gesetz hier von großer Bedeutung.

Fällt in der Software eine Prognose für eine Person schlechter aus, als für eine andere, darf diese nicht ausschließlich von den oben genannten Eigenschaften abhängig sein. Ein Anwendungsbeispiel ist die von dem Unternehmen Amazon entwickelte Software zur Bewertung von Bewerbern (Vincent, 2018). Die Entwicklung des Programms wurde eingestellt, weil es die Bewerber aufgrund des Geschlechts unterschiedlich bewertet hat.

Digitale Entwicklungen schreiten schnell voran und bringen neue Herausforderungen. Daher hat sich der Bundesminister der Justiz und für Verbraucherschutz 2017 für ein digitales Antidiskriminierungsgesetz und mehr Transparenz bei Algorithmen ausgesprochen. Er forderte den Aufbau einer Digitalagentur, die Algorithmen, sowie „das Internet der Dinge und das Leben in der digitalen Welt“ regelt (Beuth, 2017).

Im November 2018 wurde daraufhin von der Bundesregierung eine Strategie entwickelt, die den Einsatz neuartiger Technologien „ethisch, rechtlich, kulturell und institutionell“ regeln soll (Die Bundesregierung, 2018, p.

4). Dabei sind die Werte, die in allen Phasen der Softwareentwicklung beachtet werden müssen „Transparenz, Nachvollziehbarkeit, Diskriminierungsfreiheit und Überprüfbarkeit“ (Die Bundesregierung, 2018, p. 39). Software aus Industrieländern kann diskriminierend für den Einsatz in Entwicklungsländern sein. Die Ursache kann hier in unvollständigen oder fehlerhaften Trainingsdaten liegen oder in politischen Organisationsmaßnahmen und Beschränkungen (Die Bundesregierung, 2018, p. 43 f.).

Eine von der Europäischen Union eingesetzte Expertengruppe für Künstliche Intelligenz fordert für die ethische Vertretbarkeit unter anderem „Fairness“. Vorteile und Kosten sollten nicht ungleich verteilt werden; gleichzeitig sollen „Personen und Gruppen vor unfairer Verzerrung, Diskriminierung und Stigmatisierung geschützt werden“, sodass Chancengleichheit besteht (Europäische Kommission, 2018, p. 14 f.). Der Kern der Leitlinien besteht aus sieben ethischen Grundsätzen. Zwar weisen nicht alle Anforderungen direkt auf Diskriminierung hin, jedoch stehen die einzelnen Punkte in engem Zusammenhang. Abbildung 3 zeigt, dass die Anforderungen untereinander verbunden sind, was die enge Wechselbeziehung hervorhebt. Die Faktoren werden im Folgenden kurz vorgestellt.

Der Punkt „Vorrang menschlichen Handelns und menschliche Aufsicht“ unterstreicht die rein unterstützende Funktion von KI-Systemen. Sie dürfen den Menschen nicht vor vollendete Tatsachen stellen, da gezielt manipulierende und unbemerkt wirkende Prozesse die Autonomie des Menschen bedrohen können. Stattdessen sollte der Mensch interaktiv in den Entscheidungsprozess eingebunden werden.

Um Vertrauen in eine Software setzen zu können, muss diese technisch robust und sicher sein. Das heißt, dass Angriffe und Sicherheitsverletzungen verhindert und Maßnahmen zur Schadensverhütung getroffen werden sollten. Ausgaben des Modells müssen präzise und dem Sachverhalt entsprechend korrekt und zuverlässig sein.

Der Grundsatz „Datenschutz und Datenqualitätsmanagement“ verdeutlicht den großen Einfluss der Daten auf die Gerechtigkeit der Ausgabe. Die Europäische Kommission fordert hier den Schutz der Privatsphäre, weil diese Aufschluss über persönliche Merkmale eines Menschen geben. Da die Daten das Verhalten des Modells beeinflussen, müssen diese genau, fehlerfrei und ausgeglichen sein. Deshalb sollten sie in allen Entwicklungsschritten getestet und dokumentiert werden.

Als vierte Anforderung gilt „Transparenz“, aufgeteilt in die Komponenten „Rückverfolgbarkeit“, „Erklärbarkeit“ und „Kommunikation“.

Der Punkt „Vielfalt, Nichtdiskriminierung und Fairness“ hebt hervor, dass besonders bei der Datenerhebung Wert auf genaue Daten gelegt werden muss, die keine benachteiligenden Züge aufweisen. Eine Maßnahme dafür ist die genaue Erörterung der Anforderungen der Software, sodass die Erhebung der Daten dahingehend konkreter beaufsichtigt werden kann. Außerdem tragen Mitarbeiter aus verschiedenen Kulturen und Hintergründen zur Fairness bei. Auch der Zugang zur Software muss durch eine nutzerorientierte Gestaltung für jeden möglich gemacht werden, wobei die besonderen Bedürfnisse von Menschen mit Behinderung beachtet werden müssen. Vor, während und nach der Entwicklung sollten Betroffene daher aktiv in den Prozess einbezogen werden.

Das gesellschaftliche und ökologische Wohlergehen ist ein weiteres Merkmal und beschreibt die Förderung der Nachhaltigkeit und Umweltfreundlichkeit. Die Beeinflussung der sozialen Kompetenzen und allgemeine Auswirkungen auf die Gesellschaft müssen besonders beachtet werden. KI-Systeme beeinflussen die „Vorstellung von sozialer Handlungsfähigkeit“ und „sozialen Beziehungen und Bindungen“ (Europäische Kommission, 2018, p. 23 f.)

Das Vertrauen in ein System kann besonders durch eine Prüfung durch Außenstehende gesteigert werden („Rechenschaftspflicht“). Vorkehrungen für mögliche Rechtsmittel ermöglichen darüber hinaus im Zweifelsfall das

Ergreifen geeigneter Rechtsbehelfe, wobei die Entscheidungsträger im Prozess für die getroffenen Entscheidungen verantwortlich sind.

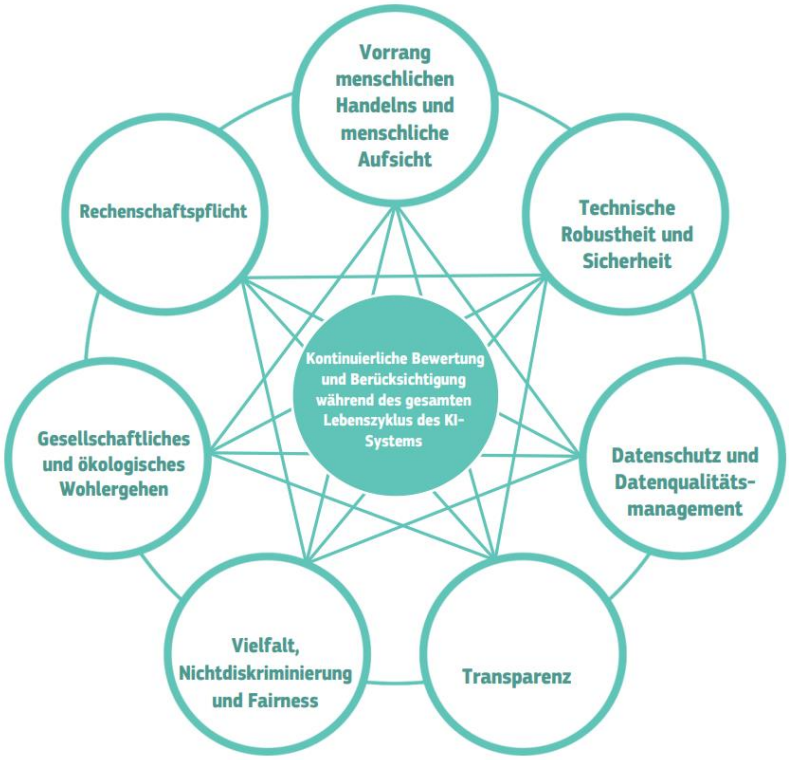

Abbildung 3: Anforderungen für vertrauenswürdige Software (Europäische Kommission, 2018)

Einen speziell auf Diskriminierung ausgerichteten Leitfaden hat die *Association for Computing Machinery* (ACM) erstellt (ACM U.S. Public Policy Council; ACM Europe Policy Committee, 2017). Die Vereinigung beschäftigt sich mit Prinzipien und Richtlinien zur verantwortungsbewussten Entwicklung von Algorithmen. Diese Prinzipien unterstützen Entwickler dabei, ethisch vertretbare Entscheidungen zu treffen. Laut ACM ist dabei das Wichtigste ein verantwortungsbewusster und transparenter Entscheidungsprozess (Association for Computing Machinery (ACM), 2018). Die im Leitfaden betrachteten Prinzipien sollen in allen Phasen der Software Entwicklung beachtet werden. Automatisierte Entscheidungen sollen den gleichen Prinzipien wie manuell getroffene Entscheidungen folgen.

Die Entwicklung von fairer Software beginnt mit der Auswahl oder Schulung aller Stakeholder. Haben die Entwickler ein Bewusstsein für Fehler im Projektablauf, die zu einer benachteiligenden Behandlung verschiedener Personengruppen führen, können sie diese von vorn herein umgehen. Es geht dabei nicht nur um allgemeine Kenntnisse über Prozesse und Algorithmen, sondern um die Kenntnis über Risiken, die im Laufe des Projekts Vorurteile verursachen können.

Zusätzlich zu der Möglichkeit, Entscheidungen nachträglich in Frage zu stellen, sollte der Herausgeber des Produkts Verantwortung für die Entscheidungen der Software übernehmen, auch wenn der Entscheidungsprozess nicht immer vollständig nachvollziehbar ist.

Ein wesentlicher Einflussfaktor auf die Ausgabe des Algorithmus ist der zugrundeliegende Datensatz. Da gerade dort Ungleichheiten und Vorurteile vorliegen können, empfiehlt ACM eine detaillierte Dokumentation des Erhebungsprozesses mit Fokus auf potenzielle Benachteiligungen. Sollte im späteren Verlauf der Entwicklung oder beim Einsatz erkennbar werden, dass dessen Ergebnisse vorurteilsbehaftet sind, ist es wichtig, die Ursache zu finden. Daher sollte der Entwicklungsprozess genau dokumentiert werden. Besonders in der Testphase dienen gezielte Tests zur Untersuchung des Modells auf eine mögliche Diskriminierung. Auch hier fördert eine gute Dokumentation das Vertrauen in die Software.

Die *International Organization for Standardization* (ISO) hat 2018 ein Subkommittee für die Standardisierung von AI Systemen gegründet (Zielke, 2020). Seitdem wird der Standard „Information technology — Artificial Intelligence (AI) — Bias in AI systems and AI aided decision making“ (ISO/IEC AWI TR 24027) entwickelt (International Organization for Standardization). Sobald die Entwicklung abgeschlossen ist, wird das die eindeutige Kennzeichnung fairer Software ermöglichen.

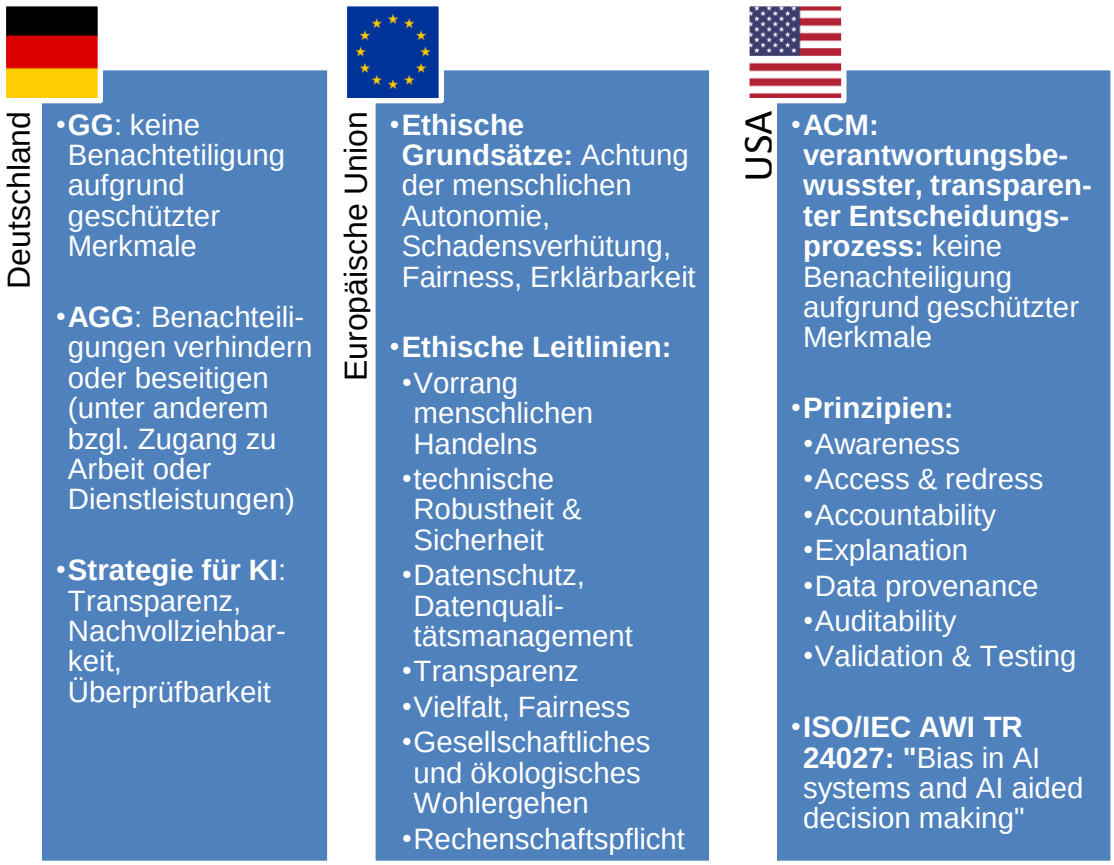

Abbildung 4: gesetzliche und ethische Standards aus Sicht Deutschlands, der EU und den USA

2.4. Ursachen für Diskriminierung und Folgen

Obwohl es zum Thema Diskriminierung klare Gesetze, sowie ethische Richt- und Leitlinien von anerkannten Organisationen gibt, tauchen häufig Applikationen auf, die diese Richtlinien nicht erfüllen. Im Folgenden wird deshalb untersucht, welche Probleme zu Diskriminierung beitragen und wie sie besonders im Zusammenhang mit Machine Learning auf das Ergebnis einwirken.

Eine Übersicht über die einzelnen Punkte, die im Folgenden untersucht werden, ist Tabelle 1 zu entnehmen. Die Tabelle gliedert sich in drei Teile,

die den Entwicklungsphasen Vorbereitung, Entwicklung und Nachbereitung entsprechen. Die einzelnen Schritte basieren auf dem Prozess nach (Suresh & Guttag, 2020) und werden in der Zeile „Aktivität“ genauer beschrieben. Analog dazu werden die dabei auftretenden Risiken für die Verminderung der Fairness aufgeführt.

In der Vorbereitungsphase geht es um die Erarbeitung des Sachzusammenhangs und die Planung der Umsetzung. Unter Entwicklung fällt sowohl die Einholung der Daten, als auch die Implementierung des Machine Learning Algorithmus. Anschließend wird das Modell in eine Software eingebettet und fertiggestellt. Jede dieser Aktivitäten birgt Risiken, die die Objektivität des Modells beeinflussen. Diese werden im Folgenden erarbeitet, um speziell die Wirkung auf ein neuronales Netz zu analysieren.

2.4.1. Datengenerierung

Die Entwicklung von KI-Systemen basiert seit einiger Zeit nicht mehr auf der Basis von Fachwissen („knowledge-based“), sondern wird durch Daten gesteuert („data-driven“) (Zielke, 2020). Aus diesem Grund ist besonders die Qualität dieser Daten für die Qualität der Ausgabe wegweisend.

So bilden die Trainingsdaten, insbesondere deren Labels die Grundwahrheiten des Algorithmus und repräsentieren die Realität (Barocas & Selbst, 2016). Sind sie nicht neutral gegenüber allen Personengruppen, kann der Algorithmus auch nicht neutral entscheiden. Wenn sie auf Situationen basieren, in denen bereits Vorurteile bestehen und praktiziert werden, dann wird sich diese Einstellung als Grundwahrheit festigen und zukünftige maschinelle Entscheidungen beeinflussen.

Ein Beispiel dafür ist die eingangs angeführte Software zur Bewertung von Bewerbern bei Amazon. Für das Training des Modells wurden Daten aus der Vergangenheit über die Einstellungen von Bewerbern verwendet, bei denen Männer jedoch bevorzugt wurden (Vincent, 2018). Dieses Verhal-

ten hat der Algorithmus erkannt und übernommen, was zu einer Benachteiligung von weiblichen Bewerbern führte.

Obwohl die Daten also inhaltlich der Realität entsprechen, kann diese das Modell negativ beeinflussen (Suresh & Guttag, 2020). Hat eine Person aus der Gruppe X systematisch ein schlechteres Ergebnis als andere, erkennt der Algorithmus daraus einen Zusammenhang und interpretiert die Zugehörigkeit zu Gruppe X als negativen Indikator. Handelt es sich bei dieser Gruppe um eine geschützte Eigenschaft (z.B. Alter, Herkunft), entsteht dadurch Diskriminierung (Calders & Zliobaite, 2013).

Das Training neuronaler Netze erfolgt zunächst durch die initiale Belegung der Gewichte, die bestimmen, wie stark ein Feature in die Ausgabe einfließt. Anhand der Trainingsdaten wird die Abweichung der Prognose vom tatsächlichen Label mit jedem Durchlauf minimiert, indem die Gewichte nachjustiert werden. Wenn die Trainingsdaten verzerrt sind und diskriminierend verteilte Label enthalten, erfolgt die Belegung der Gewichtungen ebenso auf diskriminierende Weise, was zu unfairen Ergebnissen führt.

Das gilt auch für Verfahren, bei denen das Verhalten von Nutzern ausgewertet wird. Diese Analyse kann beispielsweise für das Anzeigen personalisierter Werbeanzeigen dienen. Dabei sollen die angeklickten Themen, die Dauer des Verweilens auf einer Seite oder andere Merkmale dazu verwendet werden, die Interessen des Nutzers auszumachen und entsprechende Werbung anzuzeigen. Das Verhalten des Nutzers ist jedoch durch persönliche Vorurteile geprägt und beeinflusst damit den Algorithmus (Barocas & Selbst, 2016).

Darüber hinaus beeinflussen traumatisierende Erlebnisse die Daten. Solche und ähnliche Ausreißer führen zu Verbindungen zwischen Attributen, die sich negativ auswirken und eine Benachteiligung verursachen (Baer, 2019, p. 73).

2.4.2. Auswahl der Population

Da nicht immer die gesamte betroffene Bevölkerung in einem Datensatz abgebildet werden kann, muss eine kleinere Population ausgewählt werden. Diese sollte eine kleine Kopie der gesamten Menge sein (Calders & Zliobaite, 2013). Die Modellierung dieser Teilmenge ist nicht trivial, denn es kommt besonders auf die Generierungsmethode der Daten an. Findet eine Befragung beispielsweise online statt, schließt das automatisch als Befragte diejenigen aus, die über keinen Internetzugang verfügen. Auch andere Dienste oder Produkte werden von verschiedenen Personengruppen weniger häufig oder gar nicht benutzt, sodass daraus stammende Daten nicht alle Gruppen repräsentieren (Orwat, 2019, p. 81).

Leicht können auch geografische Regionen unterrepräsentiert sein, sodass der Algorithmus keine ausgeglichene Ansicht über alle Regionen erhält. Je nach Anwendungsszenario liefert der Algorithmus für Menschen aus dieser Region dann schlechtere Ergebnisse. Gleiches gilt für veraltete Daten, denn diese repräsentieren aktuelle Umstände nicht ausreichend (Suresh & Guttag, 2020).

Ein typisches Beispiel, das zeigt, dass ein nicht repräsentativer Datensatz oft unbemerkt generiert wird, ist die Kontrolle von Arbeitnehmern. Wird eine bestimmte Abteilung häufiger kontrolliert als andere, werden dort häufiger Fehler entdeckt. Diese Abteilung wird im Datensatz zur Fehlerdokumentation daher übermäßig stark repräsentiert. Die Wahrscheinlichkeit für das Auftreten eines Fehlers eines Arbeitnehmers aus dieser Abteilung fällt höher aus, als in der Vergleichsgruppe ohne häufige Kontrollen. Ein Algorithmus prognostiziert daher für Mitarbeitende aus dieser Abteilung eine höhere Fehlerrate. Das Vorurteil verstärkt sich sogar, wenn die Abteilung daraufhin noch häufiger kontrolliert wird (Barocas & Selbst, 2016).

Neben der ausgeglichenen Repräsentation der ausgewählten Bevölkerung ist es auch problematisch, wenn diese nicht die Population darstellt, die letztendlich von dem Ergebnis des Algorithmus betroffen ist. Das Problem

wird verstärkt, wenn das Modell genau auf diese Daten hin optimiert wird (wenn also auch die Evaluierungsdaten mangelhaft sind), weil das eine bessere Qualität und somit gegebenenfalls ökonomische Vorteile impliziert (Suresh & Guttag, 2020).

Ein Anwendungsbeispiel ist eine Gesichtserkennungs-Software, die für dunkelhäutige Frauen ein ungenaueres Ergebnis lieferte (Suresh & Guttag, 2020). Die Ursache war, dass sie seltener in den Trainingsdaten vorkamen. Dieses Problem verstärkte sich, als für die Evaluierung ein Datensatz verwendet wurde, der weniger als 10% dunkelhäutige Frauen enthielt. Dadurch fällt die ungenügende Qualität des Modells kaum auf. Das Ergebnis der Software wirkt sich nachteilig auf die unterrepräsentierten Gruppen aus und ist in vielen Fällen unpräzise (Wang, et al., 2019).

2.4.3. Festlegung der Features und Labels

Bevor aus der ausgewählten Population Informationen zum Training des Modells generiert werden, müssen die Labels festgelegt werden. Soll berechnet werden, ob sich ein Kunde für ein bestimmtes Kreditkartenprogramm eignet, kann der Entwickler die Kategorien „geeignet“ und „nicht geeignet“ festlegen. Der Algorithmus bestimmt, welche Kunden als riskant gelten und vergibt das entsprechende Label. Dem zugrunde liegt die Hypothese „Riskante Kunden verhindern in diesem Förderprogramm den Profit“. In diesem von (Baer, 2019, p. 60 f.) betrachteten Fall wurde jedoch schnell klar, dass diese Aussage mit Vorurteilen behaftet ist. Die als riskant eingestuften Kunden zahlten hier mehr Zinsen und Gebühren, während nicht riskante Kunden das Programm nicht genügend nutzten.

Dieses Beispiel zeigt, dass die Software auf einer objektiven Hypothese beruhen muss. Das Problem wird aufgedeckt, indem das entwickelte Modell und der Gesamtkontext für die Nutzenden erklärbar gemacht werden. Dadurch ist klar, auf welchen Annahmen die Ergebnisse beruhen, sodass das Vertrauen in die Zuverlässigkeit des Modells gestärkt wird. Falls die

zugrundeliegenden Annahmen fehlerhaft oder vorurteilsbehaftet sind, kann das Problem leichter entdeckt und verhindert oder behoben werden.

In vielen Fällen können die Label eindeutig definiert werden, zum Beispiel wenn eine E-Mail in die Kategorien „Spam“ oder „kein Spam“ eingeordnet werden soll. Dafür gibt es eindeutige Indikatoren, die bei der Entscheidung berücksichtigt werden und ein klares Ergebnis liefern. Die Definition der Label ist jedoch nicht immer so trivial. Viele Zielvariablen sind nicht eindeutig erfüllbar, weil die Auswahl der Features, die zur Entscheidung beitragen, variabel ist. So ordnen verschiedene Personen einem Datenobjekt ggf. unterschiedliche Label zu.

Ein Beispiel dafür ist das eingangs erwähnte COMPAS System (Orwat, 2019). Es schätzt das Risiko dafür ein, dass ein Straftäter in der Zukunft erneut straffällig wird. Den betrachteten Personen wird das Label „wird zukünftig wieder straffällig werden“ bzw. „bleibt zukünftig straffrei“ zugeteilt. Ob dieses Label korrekt ist, kann zum Bewertungszeitpunkt nicht abschließend verifiziert werden. Ob ein ehemaliger Straftäter den Rest seines Lebens straffrei bleibt, zeigt sich erst dann, wenn er wieder eine Straftat begeht oder sogar erst bei seinem Tod. Dadurch können Fairness und Transparenz des Modells abnehmen. Für ein Problem gibt es verschiedene Lösungswege, die von den Labels, Features und deren Gewichtung abhängig sind (Barocas & Selbst, 2016).

Neuronale Netze bestehen aus vielen gewichteten Neuronen, die jeweils die Features aus dem Datensatz repräsentieren. Insofern steht damit fest, welche Variablen in die Ausgabe des Algorithmus als Einflussfaktoren einfließen. Auch falls diese inhaltlich nichts mit der Ausgabe zu tun haben, oder gesetzlich zu tun haben sollten, wird die Maschine versuchen, einen Zusammenhang herzustellen. Aus diesem Grund ist die Auswahl der Features ausschlaggebend für ein genaueres Ergebnis und für die Herstellung korrekter, objektiver Zusammenhänge.

Eine simple, aber oberflächliche Lösung wäre das Herausnehmen geschützter Merkmale. Das Problem liegt jedoch tiefer und es ist erforderlich, die Zusammenhänge zwischen Attributen von Menschen zu betrachten. So kann beispielsweise der Wohnort eines Menschen in manchen Fällen auf die Herkunft hinweisen, wenn in dieser Gegend eine bestimmte Nationalität häufiger vertreten ist (Orwat, 2019, p. 77 f.).

Eine Variable dieser Art wird „Proxy“ genannt; sie ist kein konkret geschütztes Merkmal, stellt aber indirekt ein Merkmal für eine Personengruppe dar. Durch die Verwendung eines Proxies ist nicht eindeutig, welche Attribute genau auf die Entscheidung einwirken, da diese Variable indirekt weitere Merkmale beinhaltet (Calders & Zliobaite, 2013). Dadurch kann die zugrundeliegende Kausalkette nicht eindeutig definiert werden. Kausale Schlussfolgerungen sind jedoch eine Grundvoraussetzung für eine wissenschaftliche Arbeitsweise (Holzinger, 2018).

Das Ziel bei der Auswahl der Features ist daher eine möglichst detaillierte, umfangreiche Wahl, sodass die Realität möglichst genau abgebildet wird. Aus verschiedenen Gründen ist das aber nicht immer möglich. Eine sehr große Menge an Features würde die Performance erheblich verschlechtern (Orwat, 2019, p. 79). Bestimmte Eigenschaften von Menschen können aus datenschutzrechtlichen Gründen nicht erfragt werden (Barocas & Selbst, 2016). Werden daher weniger Attribute ausgewählt, können entscheidende Details verloren gehen, was dazu führt, dass Personengruppen nicht mehr ausreichend repräsentiert werden oder die Realität ungenau widergespiegelt wird (Orwat, 2019, p. 79).

Wird beispielsweise eine Software entwickelt, die aus Bewerbern entscheidet, wer für eine ausgeschriebene Stelle geeignet ist, ist zu analysieren, welche Eigenschaften ein Bewerber aufweisen sollte und durch welche Attribute diese gemessen werden können. Wird ein hoher Stellenwert auf die Beschäftigungsdauer gelegt, kann das eine Benachteiligung für weibliche Bewerber darstellen, da sie im Allgemeinen eine höhere Wechselrate aufweisen (Barocas & Selbst, 2016). Bei Betrachtung anderer Ge-

sichtspunkte fällt das Ergebnis anders aus. Es kommt daher darauf an, zu analysieren, welche Eigenschaften im Sachzusammenhang wirklich auf das Ergebnis einwirken. Der Entwickler muss die Anforderungen und Ziele des Unternehmens berücksichtigen (Barocas & Selbst, 2016).

Zudem ist die Auswahl der Features subjektiv. Geht der Entwickler aufgrund persönlicher Vorurteile davon aus, dass ein Feature keinen Einfluss auf das Ergebnis des Algorithmus nehmen sollte, wird dieses in der Datenerhebung nicht erfragt. Das kann persönliche Vorurteile des Entwicklers im System einführen (Baer, 2019, p. 61 f.).

2.4.4. Datenerhebung

Für das Trainieren und Evaluieren neuronaler Netze werden große Mengen an Daten benötigt, damit das Training möglichst präzise ablaufen kann. Um Repräsentativität gewährleisten zu können, ist es erforderlich, eine Menge an Menschen oder Institutionen damit zu beauftragen, die Daten zu erheben. Abhängig von der gewählten Methode wirken auch hier die subjektiven Einstellungen der Verantwortlichen auf den Prozess ein.

Eine weitere Ursache für Diskriminierung ist eine mindere Qualität für bestimmte Personengruppen oder wenn die Attributwerte mit verschiedenen Maßstäben festgehalten werden (Suresh & Guttag, 2020). Das kann zum Beispiel an fehlender oder unterschiedlich aufgebauter Infrastruktur liegen.

Im Laufe der derzeit weltweit kursierenden COVID-19 Pandemie werden die Infektions- und Sterblichkeitszahlen verschiedener Länder dokumentiert. Wie Prof. Dr. Dr. Tobias Kurth, Professor für Bevölkerungsgesundheit und Epidemiologie, beschreibt, sind diese Zahlen nicht immer vergleichbar, „da andere Länder ihre Zahlen anders melden. Sie testen anders, oder sie klassifizieren Patienten anders“ (Shield & Abbany, 2020). Basieren Algorithmen auf diesen Messungen, stimmt das Verhältnis zwischen den Personengruppen nicht überein, sodass ein verzerrter Datensatz entsteht.

Der aus der Befragung entstandene Datensatz muss nicht nur die betrachtete Bevölkerung ausgeglichen abbilden. Auch die Verteilung der Label muss ausgewogen sein (Baer, 2019, p. 89 f.). Bei einem Algorithmus, der einen Bewerber als „geeignet“ oder „nicht geeignet“ klassifizieren soll, ist es unmöglich, korrekte Ergebnisse zu erzielen, wenn im Trainingsdatensatz nur wenige Objekte mit dem Label „geeignet“ enthalten sind. Dem Algorithmus fehlt die Entscheidungsgrundlage für dieses Kriterium.

Ein Großteil der Ursachen für Diskriminierung liegt in der Erhebung der Daten, da sie durch kleine Fehler oder Umstände, die der Entwickler nicht berücksichtigt, unbemerkt unausgeglichen oder vorurteilsbehaftet sind. Wie bereits erwähnt, legen diese Daten jedoch die Grundwahrheiten des Modells fest, sodass diese Vorurteile oder Verzerrungen fortgesetzt und verstärkt werden. (Baer, 2019, p. 69 ff.) behandelt viele Details, die sich auf die Qualität der Daten auswirken. Dazu gehört die Tatsache, dass auch faktische Daten, beispielsweise das Einkommen, unterschiedlich angegeben werden können. Abhängig davon, wer die Daten dokumentiert und welche Interessen die Person dabei verfolgt, schätzt sie das Datum unterbewusst nach den eigenen Interessen ein. Auch der Aufbau eines Fragebogens sollte sorgfältig durchdacht werden, weil Drop-Down-Listen und das Gruppieren von Personengruppen Auswirkungen auf die Antwort des Befragten haben. An dieser Stelle können nicht alle Details eingehend aufgeführt werden, jedoch lohnt es sich, diese zu beachten.

2.4.5. Nachbearbeitung, Unterteilung

Nachdem ein Datensatz erstellt wurde, muss dieser bereinigt und so angepasst werden, dass das neuronale Netz diesen verarbeiten kann. Bei genauerer Betrachtung der Attributwerte – besonders wenn es sich um numerische Werte handelt – kann sich herausstellen, dass einige davon unbedeutend sind oder aufgrund minimaler Werte praktisch keine Bedeutung im Sachzusammenhang haben. Wie (Baer, 2019) am Beispiel von

Daten über die Kreditrückzahlung zeigt, verfälscht eine Löschung dieser scheinbar unbedeutenden Werte das Ergebnis, wenn dadurch ähnliche Daten betroffen sind, die nicht unbedeutend sind.

Darüber hinaus werden Attribute oft in Rankings also numerische Werte umgewandelt. Auch diese Bearbeitung der Daten verursacht Diskriminierung, wenn sich das Ranking in der Realität ändert, aber die Bezeichnungen, die den Werten zugeordnet sind, nicht korrigiert werden. Wird beispielsweise das Feature „besuchte Universität“ anhand eines offiziellen Rankings auf numerische Werte übertragen, so müssen die Namen der Universitäten angepasst werden, wenn sich dieses Ranking ändert. Falls das nicht geschieht, ist der Algorithmus diskriminierend gegenüber Besuchern einer bestimmten Universität, obwohl diese möglicherweise an Ansehen gewonnen hat. Der Algorithmus würde keine faire Entscheidung treffen, sondern nach veralteten, verzerrten Umständen handeln.

2.4.6. Entwicklung des Modells

Auch der gewählte Algorithmus kann Ursache für ungenaue oder benachteiligende Klassifizierungen und Prognosen sein. Dazu gehören Vorgehen, die auf Entscheidungsbäumen basieren, wie zum Beispiel Random Forests (Baer, 2019, p. 91). Dabei wird der gesamte Datensatz in Kategorien (Zweige) unterteilt. Ein Individuum mit außergewöhnlichen Eigenschaften fällt gegebenenfalls in eine Kategorie, mit deren zugehörigen Personen es nur wenige Gemeinsamkeiten hat. Trotzdem muss es die Entscheidungen für diese Gruppe mittragen, obwohl diese zwar für die Gruppe als Ganzes passend erscheinen, aber das Individuum dabei benachteiligen.

Nach der Einteilung in eine Menge von Zweigen und Blattknoten, kann es vorkommen, dass für eine Gruppe nur sehr wenige Daten übrig bleiben. Besonders wenn der Baum sehr detailliert unterteilt wird, ist das der Fall. Das macht es fast unmöglich, ein Muster zuverlässig zu erkennen und faire Entscheidungen zu treffen (Baer, 2019).

Die Trainingsdaten werden häufig über Umfragen generiert. Oft beantworten die Nutzenden eine Frage, indem sie aus einer Menge von Antworten die am ehesten zutreffende auswählen oder keine Antwort geben, falls dies möglich und gewünscht ist. Bei der Regressionsanalyse kann das zu einem Problem werden, wenn eine bestimmte Gruppe bei einer Frage häufiger keine Antwort gibt. In diesem Fall fehlt dem Algorithmus die Entscheidungsgrundlage auf Basis dieses Features.

(Fushiki & Maeda, 2020) nennen als Beispiel eine Befragung über das monatliche Einkommen. Abbildung 5 zeigt die Verteilung der Antworten und den Zusammenhang zwischen dem Alter und dem Einkommen der Befragten. Personen im Alter von 20 bis 40 Jahren gaben nur selten Auskunft über die Höhe des Einkommens. Obwohl die ausgewählte Population im Alter ausgeglichen war, lagen nicht gleich viele Antworten für alle Altersklassen vor. Dadurch repräsentiert der Datensatz die Personengruppen auf unausgeglichene Weise.

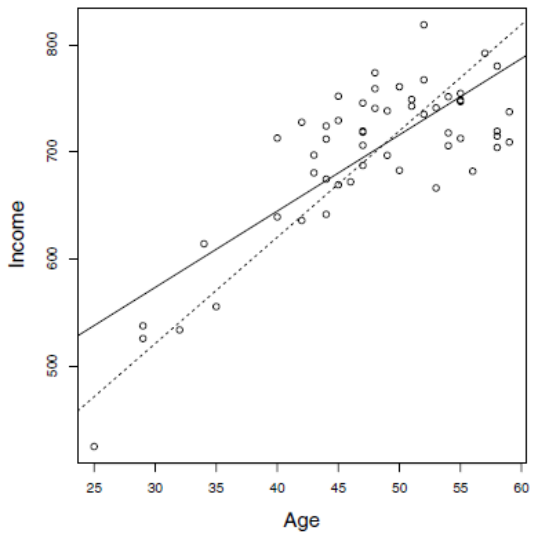

Abbildung 5: Verteilung der Attributwerte, Regressionsgeraden für die geschätzten Werte (gestrichelt) und die tatsächlichen Attributwerte (durchgezogen). Quelle: (Fushiki & Maeda, 2020)

Bei der hier verwendeten linearen Regression wird eine Regressionsgerade entwickelt, die den Zusammenhang des Labels zum Feature anhand der Methode der kleinsten Quadrate darstellt. Hier handelt es sich um den Zusammenhang zwischen Alter und Höhe des monatlichen Einkommens. Für eine schrittweise Optimierung der Geraden wird die Abweichung der tatsächlichen Werte von der Regressionsgeraden jeweils quadriert und dann aufsummiert. Diese Summe wird durch die Anpassung der Faktoren in der Funktionsgleichung bzw. der Gewichte minimiert (Windzio, 2013).

(Fushiki & Maeda, 2020) haben zwei Regressionsgeraden modelliert (Abbildung 5). Die gestrichelte Gerade gilt für die über den gesamten Datensatz geschätzten Werte, die durchgezogene Gerade für Datenobjekte, bei denen tatsächlich geantwortet wurde. Beide Geraden haben eine unterschiedliche Steigung. Außerdem setzt die Regressionsgerade für die tatsächlich Antwortenden wesentlich höher an. Das liegt daran, dass mit steigendem Alter das Gehalt steigt. Da Menschen im höheren Alter eine höhere Antwortrate hatten, nimmt der Algorithmus an, dass das Gehalt insgesamt höher ist. Schließlich ist das Ziel des Algorithmus, die Minimierung der Abweichung der Antwort von der Gerade. Das Ergebnis ist also keine realistische und faire Einschätzung, sondern es spiegelt nur die Situation derer wider, die häufiger eine Antwort gegeben haben.

2.4.7. Ausführung, Nachbearbeitung

Bei der Integration in das System, in dem der Algorithmus eingesetzt werden soll, treten Risiken für ein diskriminierendes Ergebnis auf. Auch wenn das Modell korrekt geplant und implementiert wurde, führt der Einsatz in einem Zusammenhang, für den es ursprünglich nicht entwickelt wurde, zu ungenauen Ausgaben (Suresh & Guttag, 2020). Im gesamten Datenerhebungsprozess wird hervorgehoben, wie wichtig es ist, dass die Daten eine kleine Kopie der betroffenen Population sind (Calders & Zliobaite, 2013).

Wird nun der Kontext geändert, sind diese Daten keine Repräsentation der Realität, obwohl das System daran trainiert und optimiert wurde.

Die Betrachtung des Entwicklungsprozesses neuronaler Netze und anderer Modelle zeigt, dass verzerrte Daten eine Verfestigung und Verstärkung vorhandener Benachteiligungen verursachen, da sie die Grundwahrheiten für spätere Entscheidungen festlegen. Eigenheiten verschiedener Algorithmen tragen dazu bei, dass die betrachteten Daten nicht gleichmäßig analysiert werden können. Dies wird bei der linearen Regression deutlich, wenn Personengruppen in einer Befragung keine Antwort geben.

Aufgrund der oft fehlenden Erklärbarkeit der Algorithmen ist Diskriminierung nur schwer bemerkbar. Unausgeglichene Trainingsdaten, diskriminierende Label und Features und andere Methoden können bewusst eingesetzt werden, sodass der Entwickler seine persönlichen Vorurteile einfließen lassen kann und das Ergebnis unbemerkt diskriminierend wirkt. Dieses Vorgehen wird als „Masking“ bezeichnet (Barocas & Selbst, 2016).

Um die Ursachen für Diskriminierung zu vermeiden oder im Nachhinein zu korrigieren, gibt es verschiedene Methoden. Sie werden im folgenden Kapitel behandelt.

a) Vorbereitung	1 Datengenerierung	2 Auswahl der Population	3 Festlegung Features und Labels
Aktivitäten	- Realität als Grundlage für die Daten	- Auswahl der abzubildenden Personen	- Analyse der relevanten Features - Festlegung der Label
Risiken	- Vorurteile in der Realität präsent - Vorurteilsgeprägtes Nutzerverhalten - Verfälschung des Regelfalls durch traumatisierende Erlebnisse	- Unter- / Überrepräsentation von Personengruppen, Regionen oder Zeiträumen - Population entspricht nicht der Zielpopulation	- Inkorrekte Hypothese - Label nicht eindeutig definierbar - Proxy-Variablen - Datenschutzrechtliche Hindernisse - Vorurteilsbehaftete Auswahl der Features
b) Entwicklung	4 Datenerhebung	5 Nachbearbeitung	6 Entwicklung des Modells
Aktivitäten	- Einholung der Daten durch Befragung oder ähnliche Methoden - Zusammenstellung zum Datensatz	- Aufbereitung der Daten - Unterteilung in Trainings- und Testdaten	- Definition des Modells - Optimierung anhand der Trainingsdaten - Evaluierung der Qualität
Risiken	- Personen, die die Daten erheben, handeln diskriminierend - Qualitative Unterschiede - Ungleiche Verteilung der Label	- Löschung scheinbar unbedeutender Datenobjekte - Numerisches Ranking	- Unpassende Gruppierung - Ungenügende Datenmenge - Fehlende Antworten bei Umfragen
c) Nachbearbeitung	7 Ausführung		
Aktivitäten	- Ausführung des Modells - Nachbearbeitung - Herstellung der Interpretierbarkeit - Einsatz im Sachkontext		
Risiken	- Einsatz in anderem Kontext		

Tabelle 1: Risiken für Diskriminierung im Laufe der Entwicklung

2.5. Methoden zur Vermeidung und Eliminierung von Diskriminierung

Damit ein System fair „handelt“ und Personen aller Art gleich behandelt, wurden einige Methoden entwickelt, die Diskriminierung verhindern, verringern oder eliminieren. Fairness ist eine Messgröße, die für die Qualität der Software aussagekräftig ist. Deshalb ist es erforderlich, durch Experimente oder Simulationen zu untersuchen, nach welchen Kriterien Entscheidungen getroffen werden, welchen Einfluss die Daten darauf haben und wie sich diese Entscheidungen auf betroffene Personengruppen auswirken (Orwat, 2019, p. 100). Anhand des Ergebnisses kann dann entschieden werden, wie das Problem gelöst werden muss.

(Friedler, et al., 2019) teilen die verfügbaren Lösungen in drei Kategorien ein: Vorverarbeitung, Modifikation des Algorithmus und Nachbearbeitung. Zu jeder dieser Kategorien werden in diesem Abschnitt Möglichkeiten vorgestellt. Zuvor werden außerdem nichttechnische Maßnahmen präsentiert.

2.5.1. Nichttechnische Maßnahmen

Entwickler und Entscheidungsträger nehmen durch ihre Handlungen starken Einfluss auf die Qualität einer Software. An vielen Ursachen für Diskriminierung ist deutlich geworden, dass durch menschliche Fehlschlüsse Fehler oder Unausgeglichenheiten in den Trainings- und Evaluierungsdaten entstehen können. Deshalb ist der erste Schritt zur Vermeidung dieses Problems die Sensibilisierung der Stakeholder. Dadurch können „rechtlich-ethische Standpunkte mit technischen Standpunkten“ besser vereint werden (Orwat, 2019, p. 132). Der Problemstellung muss ausreichend Aufmerksamkeit geschenkt werden, sodass die Entwickler von vorn herein im Hinblick auf Diskriminierung sensibler arbeiten (Friedman & Nissenbaum, 1996, p. 345). Die Expertengruppe der Europäischen Union für künstliche Intelligenz empfiehlt daher, das Entwicklerteam aus Menschen mit unter-

schiedlichen Hintergründen (z.B. Kultur, Alter, beruflicher Hintergrund) zusammenzusetzen (Europäische Kommission, 2018, p. 29).

Da die Trainingsdaten sehr umfangreich sind und oftmals in vielen verschiedenen Regionen oder Ländern erhoben werden, ist es oft erforderlich, eine Menge von Personen oder Institutionen damit zu beauftragen, die Daten zu erheben. Da jede von ihnen eigene Vorstellungen und Vorurteile hat, ist es nach (Baer, 2019, p. 168) von hoher Bedeutung, diesen Prozess genau zu definieren und zu standardisieren. Die einzelnen Arbeitsschritte sollten effizient und nicht zu umfangreich sein, damit das Vorgehen möglichst einheitlich ist. Optimal sind zehn bis 25 Schritte. Dabei sollten besondere Vorkehrungen getroffen werden, die verhindern, dass menschliche Vorurteile in den Prozess einfließen. Darüber ermöglicht eine geringere Anzahl an Personen, die die Daten erheben, eine gezieltere Beaufsichtigung und Anleitung des Teams.

Auch die Auswahl der Features, die in das neuronale Netz einfließen, ist als kritisch zu betrachten. Anstatt Features auszuwählen, die nach Meinung des Entwicklers einen Einfluss auf das Ergebnis haben, können sie zum Beispiel zufällig ausgewählt werden, um persönliche Vorurteile zu vernachlässigen (Baer, 2019, p. 168).

Sollten die Einflussvariablen doch manuell ausgewählt werden, ist es empfehlenswert diese schriftlich zu dokumentieren. Für menschliche, im Entwicklungsprozess getroffene Entscheidungen ist eine ausführliche, begründete Dokumentation von großer Bedeutung (Orwat, 2019, p. 100). So ist es im Fall einer Diskriminierung leichter, die Ursache zu identifizieren und zu beheben.

2.5.2. Vorverarbeitung

Falls die Trainingsdaten trotz allen Maßnahmen mangelhaft sind, müssen diese Daten bearbeitet werden, bevor sie vom Algorithmus für das Trai-

ning verwendet werden. (Kamiran, et al., 2013) stellen dazu zwei Möglichkeiten vor: Die gezielte Änderung ausgewählter Label und die Änderung der Gewichtung. Diese beiden Ansätze werden im Folgenden erläutert.

Bei Daten, die aus Umgebungen stammen, in denen Personen bereits in der Realität systematisch benachteiligt werden, sind vor allem die vergebenen Label zweifelhaft. Aus diesem Grund sollen diese gezielt so geändert werden, dass bei der bevorzugten Personengruppe positive Label in negative geändert werden. Gleichzeitig findet bei der benachteiligten Gruppe eine Umwandlung der negativen Label in positive statt. Es gilt dabei zu beachten, dass das Verhältnis der Label erhalten bleiben muss. Wird also ein negatives Label in ein positives verändert, muss gleichzeitig ein positives zu einem negativen geändert werden, sodass es hinterher die gleiche Anzahl positiver und negativer Klassen gibt, wie vorher.

Um die zu korrigierenden Label zu identifizieren, muss das Klassifizierungs-Modell einmal trainiert werden, sodass die Ausgabe für jedes Objekt im Datensatz bekannt ist (meist ein Wert zwischen 0 und 1). Anhand dieses Wertes wird nun ausgewählt, welches Datenobjekt ein neues Label bekommt. Dazu werden die Ergebnisse nach dem geschützten Merkmal sortiert, das in diesem Fall fälschlicherweise der Indikator für die Zuordnung zu einer Klasse ist (z.B. das Geschlecht). Aus der benachteiligten Personengruppe werden Objekte ausgewählt, die zur negativen Klasse gehören und von dem Modell den höchsten Wert erhalten haben. Dabei wird angenommen, dass eine höhere Ausgabe die Zuordnung zu einer positiven Klasse bewirkt. Diese Objekte erhalten nun ein positives Label. Analog werden bei Objekten aus der bevorzugten Gruppe mit positiver Klassifizierung diejenigen mit einem positiven Label und den niedrigsten Ausgaben zur Änderung der Klasse ausgewählt. So bleibt die Anzahl der einzelnen Klassen gleich, aber die Verteilung der Klassen ist nicht streng an ein geschütztes Merkmal gebunden.

Bei der zweiten Maßnahme von (Kamiran, et al., 2013) wird anhand berechneter Gewichte ein zweiter ausgeglichener Datensatz erstellt. Objekte

aus der benachteiligten Gruppe mit einem positiven Label erhalten ein höheres Gewicht als diejenigen mit einem negativen Label. Genauso erhalten Objekte aus der bevorzugten Gruppe mit einem positiven Label ein niedrigeres Gewicht, als diejenigen mit einem negativen Label.

Für die Berechnung der einzelnen Gewichte *w* wird die in einem fairen Kontext erwartete Wahrscheinlichkeit P_{exp} berechnet, mit der ein Objekt ein geschütztes Merkmal A und ein positives bzw. negatives Label L besitzt. Dieser Wert wird durch die Wahrscheinlichkeit P_{obs} geteilt, mit der das Objekt tatsächlich diese Eigenschaften besitzt.

Es ergibt sich die Gleichung

$$w = \frac{P_{exp}(A) \times P_{exp}(L)}{P_{obs}(A) \times P_{obs}(L)}$$

Handelt es sich beispielsweise bei dem geschützten Merkmal um das Geschlecht und bei dem Label um + oder -, kann das Gewicht für ein Objekt aus der Gruppe weiblich (f) und der positiven Klasse (+) folgendermaßen berechnet werden:

$$w = \frac{P_{exp}(f) \times P_{exp}(+)}{P_{obs}(f) \times P_{obs}(+)}$$

Diese Gewichte werden dem Datensatz hinzugefügt und das Modell wird anhand dessen trainiert. (Kamiran, et al., 2013) zeigen, dass ein solcher Datensatz ausgeglichen und keine Diskriminierung vorhanden ist.

Falls ein Algorithmus mit diesen Gewichten nicht umgehen kann, wird stattdessen ein weiterer Datensatz erstellt, für den einzelne Objekte aus dem ursprünglichen Datensatz herausgenommen werden. Die Gewichte dienen dann als Wahrscheinlichkeit, mit der ein Datum dafür ausgewählt wird. Das führt dazu, dass einige Objekte doppelt vorkommen oder entfallen. Das Training auf Basis dieses neuen Datensatzes ergibt ein faires Ergebnis.

2.5.3. Modifikation des Algorithmus

Wie in Kapitel 2.4.6 unter dem Punkt „Definition des Modells“ erläutert wurde, sind Fragen, bei denen viele Personen keine Antwort gegeben haben, besonders dann problematisch, wenn die Methode der kleinsten Quadrate für die lineare Regression verwendet wird. (Fushiki & Maeda, 2020) nennen in ihrer Arbeit eine Lösung für dieses Problem.

Zunächst muss analysiert werden, wie sich die Variablen gegenseitig beeinflussen. Soll der Algorithmus die Variable X (z.B. Einkommen) prognostizieren, muss untersucht werden, von welchem Feature diese Variable und die Eigenschaft „Antwort gegeben“ abhängig sind. Sollte hier ein Zusammenhang erkennbar sein (z.B. zum Alter des Befragten), wird daraus ein Kausalitätsdiagramm erstellt. Im nächsten Schritt wird eine weitere Variable U identifiziert, die mit der gefragten Variable X in Zusammenhang gebracht werden kann. Auf Basis dieser Variable U wird dann die gewichtete Methode der kleinsten Quadrate berechnet.

2.5.4. Nachbearbeitung

Die Möglichkeit zur Nachbearbeitung einer diskriminierenden Klassifizierung oder Prognose kann an dem Beispiel von Entscheidungsbäumen verdeutlicht werden (Kamiran, et al., 2013). Bei diesem Vorgehen werden die verschiedenen Profile der Datenobjekte in echte Teilmengen aufgeteilt. So kann ein neues Objekt mit unbekanntem Label eindeutig einem Blattknoten zugeordnet werden. Es erhält dann das Label, das am häufigsten in dieser Gruppe oder Region vorkommt.

Das Verfahren ist dann diskriminierend, wenn die Profile sehr homogen sind, das heißt in einer Gruppe befinden sich mehr Personen aus der benachteiligten Gruppe. Ein neues Objekt, das dieser Gruppe zugeordnet wird, erhält dann auch dieses benachteiligende Label und wird somit sys-

tematisch diskriminiert. Falls gegen diese Problematik keine Gegenmaßnahmen eingeleitet werden, kann sich diese Diskriminierung verfestigen.

Eine Möglichkeit dem entgegenzuwirken ist die gezielte Änderung der vergebenen Label (Kamiran, et al., 2013). Dafür werden die Genauigkeit und der Diskriminierungsgrad des Modells nachverfolgt. Folgende Gleichung dient der Bemessung der Diskriminierung im Datensatz D, für das geschützte Merkmal S, die benachteiligte Gruppe f und die bevorzugte Gruppe m:

$$disc_{S=f}(D)\frac{|\{X \in D|X(S) = m, X(Class) = +\}|}{|\{X \in D|X(S) = m\}|} - \frac{|\{X \in D|X(S) = f, X(Class) = +\}|}{|\{X \in D|X(S) = f\}|}$$

In dieser Gleichung wird die Wahrscheinlichkeit berechnet, dass ein Datenobjekt zur bevorzugten Gruppe m gehört und ein positives Label $+$ erhält. Davon wird die Wahrscheinlichkeit subtrahiert, mit der eine Person zu der benachteiligten Gruppe f gehört und die positive Klasse $+$ hat. Es geht also um den Unterschied in der Wahrscheinlichkeit, mit der die bevorzugte, bzw. benachteiligte Person positiv bewertet wird. Je größer diese Differenz ist, desto größer ist der Diskriminierungsgrad im Datensatz.

Die Label aus dem fraglichen Blattknoten werden dann so geändert, dass die Diskriminierung möglichst stark sinkt, aber nur wenige Abstriche bei der Genauigkeit gemacht werden müssen. Dadurch kann es vorkommen, dass Blattknoten zusammengeführt werden. Das Ergebnis ist ein fairer und gleichzeitig präziser Entscheidungsbaum.

3. Erklärbarkeit von Algorithmen

Nach Karl Poppers hypothetisch-deduktivem Modell wird die Grundlage einer Wissenschaft durch eine kausale Erklärung gebildet, „um Sachverhalte aus Gesetzen und Bedingungen deduktiv abzuleiten“ (Holzinger, 2018). Machine Learning Modelle ziehen Schlussfolgerungen und arbeiten sehr präzise. Gleichzeitig gibt es neben Diskriminierung einen weiteren großen Nachteil: Die Modelle sind häufig undurchsichtig, das heißt es ist nicht erkennbar, auf welcher Grundlage die Prognose inhaltlich basiert. Auch wenn die Ausgabe korrekt ist, muss diese gemäß dem Modell von Popper durch das Aufzeigen der Kausalität belegt werden, besonders wenn das Modell im wissenschaftlichen Kontext eingesetzt wird.

Dabei gilt anzumerken, dass es bei „Erklärbarkeit der Algorithmen“ nicht darum geht, die Funktionsweise des verwendeten Algorithmus im Detail zu erklären. Vielmehr ist die inhaltliche Vorgehensweise des Modells gemeint, um nachzuvollziehen, nach welchen Kriterien der Algorithmus entscheidet. Diese Kriterien sind für jedes Modell unterschiedlich und daher nicht allgemeingültig, wie bei der Funktionsweise des Algorithmus.

„Explainable AI“ ist ein zentrales Thema in der künstlichen Intelligenz. In diesem Teil der Arbeit werden die aktuellen Erkenntnisse aus der Fachliteratur erarbeitet. Nach einer kurzen Definition des Begriffs „Erklärbarkeit“ steckt Kapitel 3.2 durch rechtliche und ethische Standards die Voraussetzungen ab, die erfüllt werden sollen und müssen. Unter dem Punkt „Ursachen für fehlende Erklärbarkeit“ werden einzelne Modelle im Detail untersucht, um darzustellen, warum deren Ausgaben schwer nachvollziehbar sind. Darüber hinaus werden die kognitiven Voraussetzungen des Menschen beleuchtet, damit Erklärungen daran ausgerichtet werden können. Im Anschluss wird dargestellt, welche Vorteile es bringt, wenn Machine Learning Modelle transparent sind. Methoden zur Verbesserung der Transparenz werden in Kapitel 3.5 vorgestellt.

3.1. Definition des Begriffs „Erklärbarkeit“

Zusammen mit der Entwicklung maschinellen Lernens hat sich der Begriff „Explainable AI“ (kurz: XAI, deutsch: erklärbare künstliche Intelligenz) entwickelt. Er wurde das erste Mal im Jahr 2004 von Van Lent genannt (Adadi & Berrada, 2018) und beschreibt das Problem, dass nicht nachvollziehbar ist, nach welchen Strukturen oder Prinzipien Algorithmen Entscheidungen treffen, was auch als „Blackbox“-Modell bezeichnet werden kann (Holzinger, 2018). Bei den Anfängen künstlicher Intelligenz war es zwar das Ziel, die Ergebnisse nachvollziehbar zu gestalten, jedoch ist dies in dem heutzutage umfangreichen Kontext sehr schwierig. Eine sehr vorteilhafte Eigenschaft der Algorithmen besteht darin, in einem abstrakten Kontext einen Zusammenhang aufzudecken. In einigen Situationen, beispielsweise in der Medizin (London, 2019), ist es jedoch notwendig, die Hintergründe für das Ergebnis nachvollziehen zu können.

Machine Learning Modelle gelten dann als erklärbar, wenn sie eine für Menschen verständliche Erläuterung liefern können, die Aufschluss über den Entscheidungsprozess oder das Hintergrundwissen gibt (Tickle, et al., 1998). Ziel ist es daher, „die Ursachen eines beobachteten Sachverhaltes durch eine sprachliche Darlegung seiner logischen und kausalen Zusammenhänge verständlich zu machen“ (Holzinger, 2018). Dadurch werden Systeme künstlicher Intelligenz nachvollziehbar, vertrauenswürdiger und effektiver (Adadi & Berrada, 2018).

3.2. Rechtliche und ethische Standards

Den gesetzlichen Rahmen für erklärbare Machine Learning Software bildet die Datenschutz-Grundverordnung, die seit dem 23.05.2018 in der Europäischen Union gilt. Sie behandelt die Verarbeitung personenbezogener Daten, denn diese muss auf nachvollziehbare Art und Weise erfolgen (DSGVO Art. 5, Abs. 1a). Falls die Daten zum automatisierten Treffen einer Entscheidung verwendet werden, ist der Verantwortliche dazu ver-

pflichtet, „aussagekräftige Informationen über die involvierte Logik sowie die Tragweite und die angestrebten Auswirkungen einer derartigen Verarbeitung für die betroffene Person“ zu geben (DSGVO Art. 13, Abs. 2, lit. f).

Laut einer Studie der Gesellschaft für Informatik ist damit die „grundsätzliche Entscheidungsstruktur“ gemeint (Gesellschaft für Informatik, 2018, p. 100). Ohne die Trainingsdaten oder den Algorithmus preiszugeben, oder spezifische Entscheidungen detailliert zu begründen, muss der Verantwortliche in der Lage sein, die Entscheidungskriterien zu nennen.

Bei KI-Systemen muss eine Begründung der Ergebnisse gegeben werden. Obwohl das nicht immer detailliert möglich ist, sollten trotzdem Methoden verwendet werden, wie beispielsweise eine transparente Kommunikation über die Fähigkeiten des Systems, oder Rückverfolgbarkeit (Europäische Kommission, 2018). Zu den bereits in Kapitel 2.3 erwähnten Anforderungen an vertrauenswürdige KI gehört auch Transparenz über die verwendeten Daten und das System. Dieser Faktor kann durch die Erläuterung der technischen Prozesse und den menschlichen Entscheidungen dabei erfolgen. Gegebenenfalls muss ein Kompromiss zwischen Erklärbarkeit und Genauigkeit getroffen werden.

Um qualitativ hochwertige Modelle zu entwickeln, setzt die Bundesregierung in Deutschland auf die Vertrauenswürdigkeit der Systeme. Dieses Vertrauen kann durch Transparenz geschaffen werden (Die Bundesregierung, 2018, p. 39). Die Verantwortlichen müssen dazu in der Lage sein, „Kriterien, Ziele und Logiken“ im Entscheidungsprozess zu nennen, damit Betroffene Einsicht darin haben können.

Zu den sieben Prinzipien für die verantwortungsbewusste Entwicklung von Algorithmen gehört das Prinzip der Erklärbarkeit (Abbildung 3). Darunter fallen Erläuterungen über die Verfahrensweise des Algorithmus und über die resultierende Entscheidung (ACM U.S. Public Policy Council; ACM Europe Policy Committee, 2017).

Davon abgesehen hat künstliche Intelligenz, die in der Medizin eingesetzt wird, eine ethische Verantwortung. Bei einer Diagnose muss eine Begründung geliefert werden, um klare Beweise zu verwenden. Eine falsche Diagnose hat fatale Folgen und ist aufgrund der fehlenden Erläuterung nicht aufdeckbar (Goebel, et al., 2018). Befragt ein Mediziner einen Experten nach dessen fachlicher Einschätzung, muss diese durch medizinische Prinzipien und Fakten belegt sein. Nur so kann der Mediziner diese Einschätzung bewerten und einordnen, um zu untersuchen, welche Faktoren und Erkenntnisse Einfluss genommen haben. In ähnlicher Weise muss ein Algorithmus, der bei medizinischen Fragen konsultiert wird, das Ergebnis begründen und sich rechtfertigen können, damit es vom Mediziner ernsthaft in Betracht gezogen werden kann (London, 2019).

Die verschiedenen Sichtweisen in Deutschland, der Europäischen Union und in den USA sind auf Abbildung 6 zusammengefasst. Bei den Richtlinien für erklärbare Software geht es vor allem darum, automatisierte Entscheidungen auf eine für Menschen verständliche Weise zu begründen und erläutern. Wie diese Erläuterungen im Detail gegeben werden, ist gesetzlich nicht vorgeschrieben. Gleichzeitig kommt es auf den Kontext an, in dem das System eingesetzt wird. Besonders im medizinischen Bereich ist es wichtig, Klassifizierungen und Prognosen hinterfragen zu können. Die Richtlinien verschiedener Organisationen haben außerdem gezeigt, dass es die Möglichkeit gibt, Kausalitäten darzustellen und die Logik des Modells zu beschreiben, sodass die Software vertrauenswürdig ist.

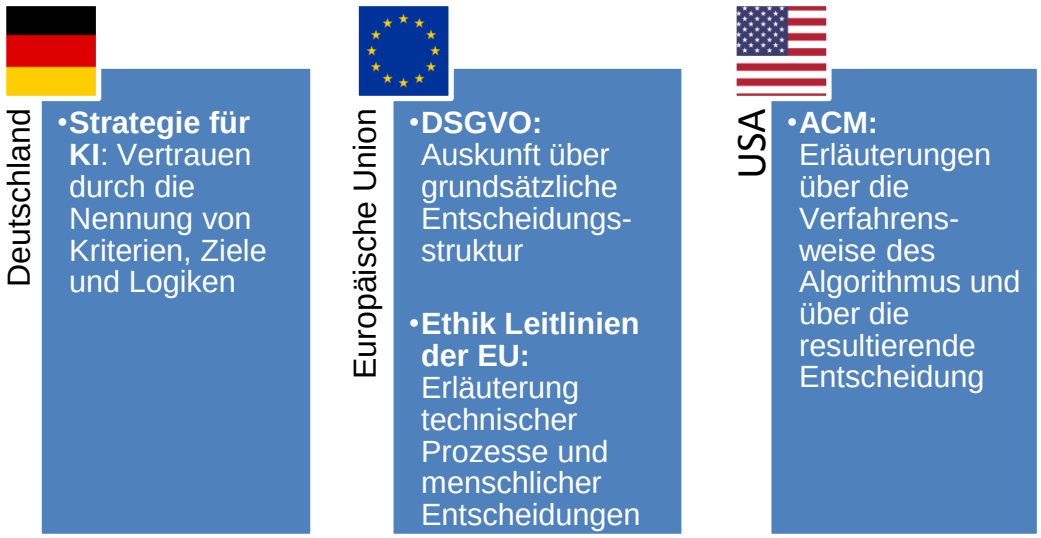

Abbildung 6: Übersicht rechtliche und ethische Leitlinien aus Deutschland, der EU und den USA

3.3. Ursachen für fehlende Erklärbarkeit

Für die Entwicklung von Machine Learning basierten Modellen gibt es verschiedene technologische Möglichkeiten. Von Entscheidungsbäumen über grafische Modelle bis hin zu neuronalen Netzen können sich die Algorithmen den betrachteten Sachverhalt selbst erarbeiten und daraufhin passende Prognosen machen. Dabei hat jedes Modell unterschiedliche Eigenschaften. Grundsätzlich gilt es abzuwägen zwischen einer guten Performance und der möglichen Erklärbarkeit des Modells (Gunning & Aha, 2019). Abbildung 7 zeigt den Zusammenhang zwischen diesen beiden Eigenschaften für verschiedene Lernstrategien. Entscheidungsbäume sind im Vergleich zu den anderen Modellen besonders gut erklärbar, weil sie verdeutlichen, anhand welcher Kriterien eine Entscheidung getroffen wird. Gleichzeitig sind diese Entscheidungen jedoch weniger zuverlässig. Im Gegensatz dazu zeugen neuronale Netze von einer sehr guten Learning Performance, das heißt die Ausgaben dieser Modelle sind sehr genau. Nachteil ist dabei die fehlende Erklärbarkeit der Ergebnisse. Grundsätzlich gilt also: Die Modelle mit der besten Performance sind diejenigen mit der geringsten Transparenz (Holzinger, 2018).

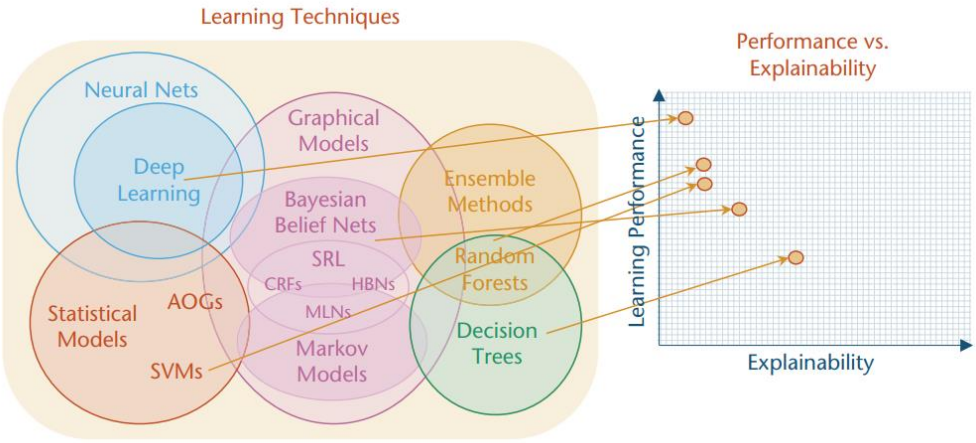

Abbildung 7: Performance vs. Erklärbarkeit verschiedener Modelle. Quelle: (Gunning & Aha, 2019)

Woran liegt es jedoch, dass die Entscheidungen neuronaler Netze so schwer nachvollziehbar sind? Systeme künstlicher Intelligenz aus den 80er Jahren wiesen dieses Problem nicht auf. Sie konnten Behauptungen mit Hilfe der Argumentationskette begründen, weil sie auf Basis der Fachkenntnisse über den Sachverhalt entwickelt wurden. Dadurch waren die Modelle zwar weniger flexibel, aber gleichzeitig sehr mächtig und vor allem erklärbar (Adadi & Berrada, 2018).

Heutige Systeme werden nicht mehr zwingend auf Basis von Fachkenntnissen entwickelt, sondern sie erarbeiten sich die zugrundeliegenden Kausalitäten selbst (London, 2019). Dieses Verfahren zeugt in den meisten Fällen zwar von einer sehr hohen Genauigkeit und Zuverlässigkeit, jedoch ist der Entscheidungsprozess aufgrund der Struktur und der Funktionsweise der Modelle nur schwer interpretierbar. Neuronale Netze bestehen aus einer generischen nicht-linearen Struktur mehrerer Schichten. Jede Schicht enthält eine Vielzahl Neuronen und viele Schichten sind verborgen (Hidden-Layer). In einigen simplen Fällen ist eine Betrachtung der Gewichtungen der Neuronen möglich, die die Features des Datensatzes darstellen. So kann die Kausalitätskette anhand der Features modelliert werden, die einen starken Einfluss auf die Prognose haben.

Bei neuronalen Netzen interagieren die Schichten nicht linear miteinander. Deshalb müssen alle Interaktionen zwischen den einzelnen Schichten in einem sehr aufwändigen Prozess genau analysiert werden, um die äußerst komplexe und verschachtelte Struktur zu durchdringen (Adadi & Berrada, 2018).

Darüber hinaus dient die Nennung der am stärksten gewichteten Features nicht unbedingt einem besseren Verständnis des Ergebnisses (Europäische Kommission, 2018, p. 27). Sie stellen nicht immer einen direkten Zusammenhang zwischen Eingabe und Ausgabe dar, denn Regelmäßigkeiten in den Daten zeigen nicht unbedingt an, wie ein Faktor im Detail auf das Ergebnis einwirkt (London, 2019).

Dieses Problem kann an einem Modell verdeutlicht werden, das die Wahrscheinlichkeit berechnet, ob ein Patient in einem Krankenhaus in der Zukunft wieder eingeliefert wird (Bayati, et al., 2014). Das Feature „Kokain-Test = negativ“ ist in diesem Modell ein Indikator dafür. Enthält der Datensatz über den betrachteten Patienten dieses Feature, steigt in dem Algorithmus die Wahrscheinlichkeit, dass er in der Zukunft wieder im Krankenhaus eingeliefert wird. Aus dieser Tatsache ist allerdings nicht erkennbar, worauf ein negatives Testergebnis hindeutet, oder welche kausalen Folgen diese Eigenschaft konkret im Hinblick auf eine Wiedereinlieferung hat. Eine genauere Untersuchung des Modells zeigte, dass das Feature nur vorhanden war, wenn der Patient den Anschein eines Drogenmissbrauchs machte. In Wirklichkeit wirkt also die Einschätzung des Arztes, ob der Patient zu einer gefährdeten Bevölkerungsgruppe gehört, auf den Entscheidungsprozess ein und weniger das Ergebnis des Kokain-Tests an sich.

Ein Feature im Datensatz enthält jedoch in manchen Fällen implizit ein weiteres Feature, das ein geschütztes Merkmal sein kann. Hier ist ein Zusammenhang zwischen der Erklärbarkeit von Modellen und der Diskriminierung erkennbar. Soll untersucht werden, ob eine automatisierte Entscheidung von einem geschützten Merkmal abhängig ist, ist anhand der Gewichtung der Neuronen bzw. der Features erkennbar, wie stark die ein-

zelnen Merkmale in Betracht gezogen werden. Sind die stärker gewichteten Features keine geschützten Merkmale, ist das Modell oberflächlich gesehen fair und erklärbar. Allerdings kann es sein, dass diese Features indirekt auf andere Beobachtungen oder geschützte Merkmale hinweisen.

Die Features der im Machine Learning verarbeiteten Datensätze sind äußerst vielfältig. Modelle wie neuronale Netze sind speziell darauf ausgelegt, diese großen Dimensionen zu verarbeiten. Die Kognitionswissenschaft zeigt jedoch, dass die Umgebung des Menschen dreidimensional gestaltet ist, sodass Datensätze der Dimension ≤ 3 menschlich sehr gut interpretierbar sind (z.B. Bilder oder Texte). In mehrdimensionalen Daten enthaltene Erkenntnisse sind für den Menschen nicht auf Anhieb extrahierbar. Zur Herstellung der Erklärbarkeit müssen die Ergebnisse deshalb in kleinere Dimensionen aufbereitet werden (Holzinger, 2018).

Die trainierten Modelle sind aus einem weiteren Grund schwer erklärbar: Für die gleichen Features und Label kann das System unterschiedliche Modelle erstellen, die nicht weniger korrekt sind. Eine Erläuterung dessen würde dann jedoch für jedes Modell verschiedene Details enthalten, was eine einheitliche automatisierte Generierung der Erklärung erschwert.

Die Ursachen für die Schwierigkeit der Schaffung von Transparenz gehen über technische Gegebenheiten hinaus. Um einen Vorgang zu erklären oder zu begründen, kann es nötig sein, vertrauliche Informationen über Personen oder das betroffene Unternehmen preiszugeben (Beining, 2019). Das würde jedoch datenschutzrechtlichen Regelungen widersprechen oder dem Unternehmen erheblich schaden, sodass eine gewisse Zurückhaltung in dieser Richtung besteht.

Darüber hinaus verursacht das Erfordernis der Nachvollziehbarkeit zusätzliche Arbeit und damit mehr Kosten – wertvolle Ressourcen in einem Unternehmen. Dementgegen sind die Vorteile, die sich durch die Transparenz ergeben, auf den ersten Blick nicht erkenn- und messbar, sodass der Anreiz fehlt oder Lösungen nachlässig erstellt werden. Es können sich

auch „Dark-Patterns“ ergeben – falsche oder irreführende Erklärungen, die ein unerwünschtes Verhalten verursachen (Beining, 2019).

Letztlich kann es schnell passieren, dass Benutzer anhand der Erläuterungen die Funktionsweise des Systems durchschauen und ihr Verhalten dementsprechend anpassen, um ein besseres Ergebnis zu erzielen (Beining, 2019). Aus diesem Grund kann eine gewisse Zurückhaltung bestehen, was die Förderung der Transparenz angeht.

3.4. Vorteile von nachvollziehbaren Entscheidungen

Trotz der Argumente gegen die Förderung der Transparenz bei KI-Systemen, gibt es einige Vorteile, die dafür sprechen, die Schwierigkeiten zu überwinden und Mühe und Zeit zu investieren, um die Transparenz der Modelle herzustellen.

Zum Beispiel sorgt Nachvollziehbarkeit für Sicherheit (Europäische Kommission, 2018, p. 26 f.). Wenn das System eine Schwachstelle in der Klassifizierung aufweist, diese aber aufgrund der fehlenden Transparenz verborgen bleibt, kann die Schwachstelle bei Angriffen leichter ausgenutzt werden. Sind die Ausgaben des Systems jedoch nachvollziehbar, werden Fehler und Schwachstellen schneller aufgedeckt, sodass es nach der Behebung zuverlässig und sicher ist.

Eine Verifikation der Prognosen ist in einigen Fachbereichen enorm wichtig. Dazu gehört die Medizin (Samek, et al., 2017). Das Kapitel über Diskriminierung im Machine Learning hat gezeigt, wie leicht sich Fehler in den Entwicklungsprozess einschleichen können, die dazu führen, dass das Ergebnis ungenau oder schlichtweg falsch ist. Eine ungenaue oder falsche Diagnose kann im medizinischen Bereich schwerwiegende Folgen haben. Deshalb muss das Ergebnis verifizierbar sein.

Die Qualität erklärbarer Modelle ist leichter optimierbar, zum Beispiel um Diskriminierung oder Verzerrungen aufzudecken und zu beheben (Samek,

et al., 2017). Die Gesellschaft für Informatik hebt in dem Zusammenhang das Problem der Hilflosigkeit des Verbrauchers hervor (Gesellschaft für Informatik, 2018). Das System muss auch für den Benutzer transparent sein, um ihm die Chance zu geben, Ergebnisse für sich selbst zu überprüfen und ggf. Beschwerde einzulegen. Diese Möglichkeit ist sogar gesetzlich gefordert (siehe Kapitel 3.2).

Aufgrund der hohen Dimension der verarbeiteten Daten ist es für den Menschen oft nicht möglich, Muster daraus zu erkennen, um inhaltlich Schlüsse aus den Daten zu ziehen (Holzinger, 2018). Die Maschine untersteht dieser Schwäche nicht und nutzt diese Fähigkeit als Methode, Regelmäßigkeiten in den Daten zu erkennen, um diese für die weitere Verarbeitung zu nutzen. Dieser Vorteil kann sich der Mensch wiederum zu Nutze machen, indem die erkannten Muster analysiert werden (Samek, et al., 2017). So kann durch Machine Learning sogar die Forschung vorangetrieben werden (Petkovic, et al., 2018).

Diese Erkenntnisse können für eine Verbesserung des Systems verwendet werden. Wenn ersichtlich ist, wie die einzelnen Features auf das Ergebnis einwirken, können Kosten und Ressourcen dadurch eingespart werden, dass irrelevante Features im Prozess gar nicht erst betrachtet werden (Petkovic, et al., 2018).

3.5. Methoden zur Herstellung der Erklärbarkeit

Zur Förderung der Transparenz des Entscheidungsprozesses gibt es nach (Wang, et al., 2019) vier Möglichkeiten. Eine ist die Lieferung einer kausalen Erklärung mit Nennung der Ursachen und Features, die für die Interpretation der Prognose relevant sind. Bei der kontrastiven Erklärung werden die zugrundeliegenden Fakten erläutert, sowie daraus resultierende Folgen. Darüber hinaus gibt es die Möglichkeit, darzustellen, welche Änderungen eine Änderung der Ausgabe verursachen. Dazu werden die Features genannt, die sich ändern müssen, damit das Datenobjekt bei-

spielsweise ein positives anstatt eines negativen Labels erhält. Das wird als „Kontrafaktische Erklärung“ bezeichnet. Ähnlich ist der Ansatz der „transfaktischen Erklärung“, bei der genauer erläutert wird, wie sich eine Änderung der Eigenschaften auf das Ergebnis auswirkt.

Eine Methode für die kontrafaktische Erklärung besteht darin, für das untersuchte Datenobjekt kontrafaktische Daten vorzustellen (Sharma, et al., 2019). Solche Daten sind so nah wie möglich am betrachteten Datenobjekt, erhalten vom Algorithmus aber eine andere Ausgabe. Damit wird deutlich, wie sich die Attributwerte ändern müssten, damit daraus das gewünschte Ergebnis resultiert.

Ein Beispiel ist die Bewertung eines Nutzers dahingehend, ob ihm ein Kredit gewährt werden sollte. Fällt das Modell ein negatives Urteil, sieht eine Erklärung folgendermaßen aus: „Wäre dein Einkommen 5.000$ höher, wäre der Kredit genehmigt worden“ (Sharma, et al., 2019). Dadurch ist klar, worauf die Bewertung basiert. Im Hinblick auf Proxy-Variablen ist es wichtig, neben der Analyse der Relevanz der Variable auch den Bezug zu anderen Features zu betrachten, um sich ein korrektes und genaues Bild zu verschaffen.

Im Folgenden beschränken sich die Beschreibungen auf die kausale Erklärung durch die Nennung der relevanten Features. Für die Identifikation dieser gibt es verschiedene Möglichkeiten, von denen im Folgenden zwei betrachtet werden. Methoden zur anschaulichen Darstellung dieser Erklärungen werden im darauffolgenden Kapitel vorgestellt.

3.5.1. Extrahieren relevanter Features aus einem neuronalen Netz

Die Besonderheit an einem Convolutional Neural Network liegt darin, dass es aus Schichten mit verschiedenen Eigenschaften besteht, die die Features der Daten erarbeiten. Mehrere Convolution-Pooling Layer sind gefolgt von einigen Fully-connected Layer. Die letzte Schicht stellt die Aus-

gabe dar (Habibi Aghdam & Jahani Heravi, 2017). Im Convolution Layer werden Filter auf die Eingabedaten angewendet, sodass gewisse Features verstärkt werden. Während des Trainings des Netzes werden die Gewichtungen der Filter so angepasst, dass die Label klar unterscheidbar sind. Abbildung 8 zeigt die Bildverarbeitung in dieser Schicht. Die Eingabe ist drei-dimensional, da das Bild aus Rot-, Grün- und Blau-Werten besteht. Darauf werden jeweils sechs verschiedene Filter angewendet, die verschiedene Eigenschaften des Bildes hervorheben, zum Beispiel durch eine Verstärkung der Kanten. Anschließend wird in einem zweiten Convolution Layer ein weiterer Filter auf die sechs Bilder angewendet.

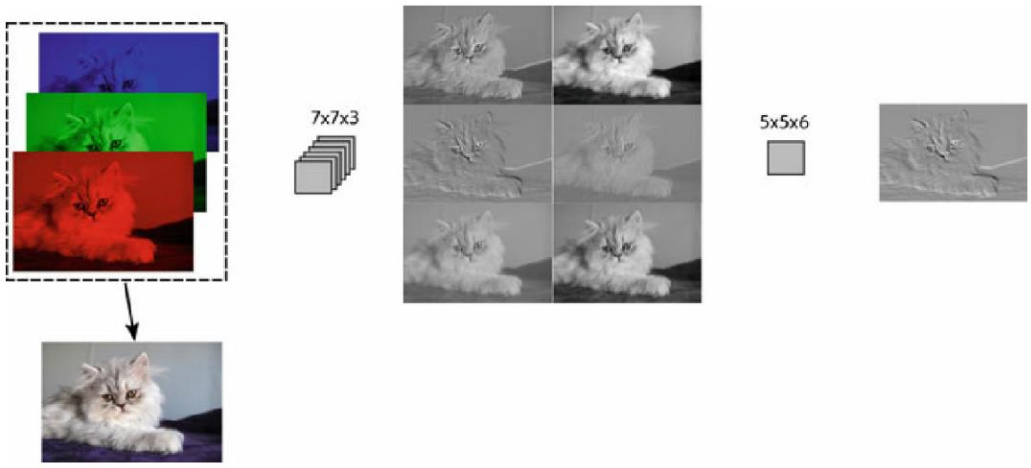

Abbildung 8: Verarbeitung einer dreidimensionalen Eingabe durch zwei Convolution Layer. Quelle: (Habibi Aghdam & Jahani Heravi, 2017)

Da die Ausgaben des Convolution Layers hoch dimensioniert sind, durchlaufen sie anschließend das Pooling Layer. Hier werden mehrere Features zusammengefasst, indem aus einer Gruppe von Werten nur der Maximal- oder Durchschnittswert weitergegeben wird. Abbildung 9 zeigt die Wirkung des Pooling Layers. Von vier aneinander liegenden Pixeln wird nur der Maximalwert 142 übernommen. Die Auswahl wird für alle 4er-Pakete vorgenommen und zu einem halb so großen Bild zusammengefügt.

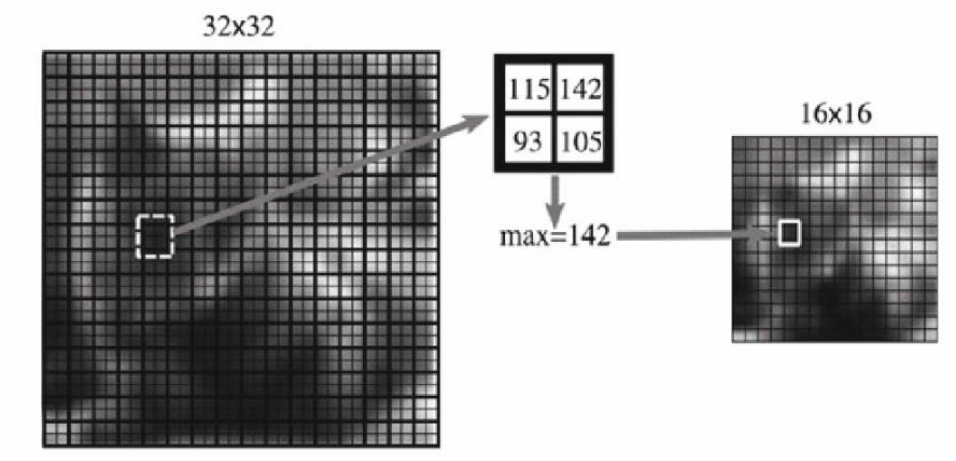

Abbildung 9: Pooling-Layer eines CNN, das die Dimensionen der Eingabe verringert. Quelle: (Habibi Aghdam & Jahani Heravi, 2017)

Durch die Verknüpfung aller Features im Fully-connected Layer erfolgt die Klassifizierung. Eine mögliche Architektur des gesamten Netzwerks ist in Abbildung 10 zu sehen. Dabei geht es um die Klassifizierung von Kommentaren im Programmcode in (non-)SATD („self-admitted technical debt“) (Ren, et al., 2019). Das Netzwerk besteht aus zwei Convolution Layers, einem Pooling Layer und einem Fully-connected Layer.

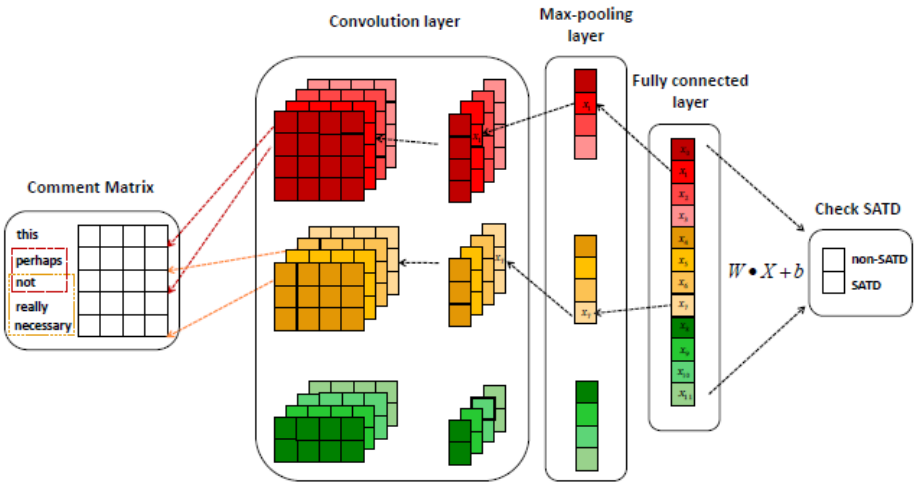

Abbildung 10: Identifizerung relevanter Features. Quelle: (Ren, et al., 2019)

An dieser Architektur wird nun erläutert, wie relevante Features nach (Ren, et al., 2019) identifiziert werden. Bei der Klassifizierung von Kommentaren soll so untersucht werden, welche Ausdrücke mit großer Wahrscheinlichkeit eine Auszeichnung als „SATD“ hervorrufen.

In Zuge dessen durchläuft ein Kommentar zunächst das gesamte trainierte Netz. Da die Features im Fully-connected Layer die Filter aus den vorherigen Schichten darstellen, wird daraus die Wahrscheinlichkeit für eine positive oder negative Einstufung berechnet. Ist die Wahrscheinlichkeit größer als 0,5, wird dieses Feature als relevant angesehen. Der zugrundeliegende Filter hat einen bestimmten Teil der Eingabe bearbeitet. In Abbildung 10 korrespondiert beispielsweise das Feature x1 mit dem rot markierten Filter, welcher die Ausdrücke „perhaps“ und „not“ verarbeitet hat. Diesen Ausdrücken wird dann die zuvor berechnete Wahrscheinlichkeit zugeordnet, mit der das Feature ein bestimmtes Label erhält. Nach Wiederholung dieses Vorgehens entsteht eine Liste von relevanten Ausdrücken, die eine positive bzw. negative Klassifizierung hervorrufen.

Ein etwas allgemeinerer Ansatz zur Berechnung der Relevanz der verwendeten Features nennt sich „Layer-Wise-Relevance Propagation“ (Samek, et al., 2017). Wie stark eine Variable das Ergebnis des Algorithmus beeinflusst, wird anhand der Aktivierung des dem Feature entsprechenden Neurons und der Verbindung mit anderen Neuronen berechnet. Um nachzuvollziehen, was das genau bedeutet, ist ein tieferer Blick in das künstliche neuronale Netz erforderlich.

Besonders die Gewichte der Neuronen, die vorher im neuronalen Netz stehen und als Eingabe für das nächste Neuron dienen, bestimmen dessen Aktivierung. Eine stärkere Aktivierung bedeutet in diesem Zusammenhang, dass das Feature, das durch dieses Neuron repräsentiert wird, einen stärkeren Einfluss auf das Ergebnis hat. Gleiches gilt für die Verbindung zweier Neuronen durch die Gewichte (Samek, et al., 2017). Zur Verbesserung der Nachvollziehbarkeit des Ergebnisses kann eine Auflistung der Features nach Relevanz sortiert erfolgen.

3.5.2. Darstellung der relevanten Features

Wie bereits beschrieben, reicht eine Auflistung der relevanten Features nicht aus, um die Entscheidung eines Machine Learning Algorithmus vollständig nachzuvollziehen (Europäische Kommission, 2018, p. 27). Es müssen weitere Maßnahmen getroffen werden, die die Erklärbarkeit verbessern und Diskriminierung aufdecken.

Geht es beispielsweise um die Klassifizierung von Bildern, kann das Ergebnis anschaulich dargestellt werden. Mithilfe eines Wärmebilds werden die relevanten Pixel hervorgehoben, sodass erkennbar ist, welche Bereiche im Bild zum Ergebnis beitragen (Goebel, et al., 2018). Relevanz besteht dann, wenn das Ergebnis bei diesem Pixel besonders empfindlich ist (Samek, et al., 2017). Abbildung 11 zeigt die Erklärung der Bildklassifizierungen, indem die Bereiche im Bild, die für die Klassifizierung am meisten betrachtet wurden, blau hervorgehoben werden.

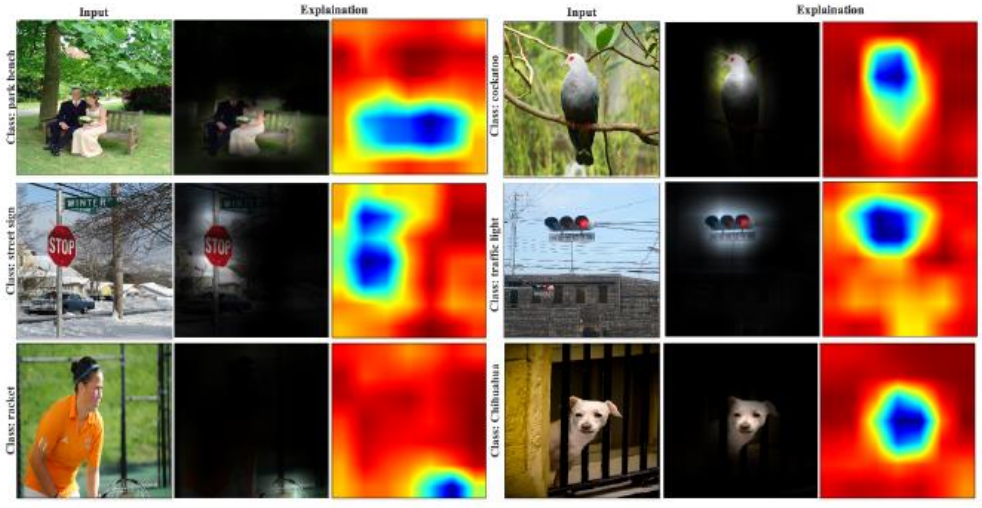

Abbildung 11: Erklärbarkeit durch Wärmebilder. Quelle: (Goebel, et al., 2018)

Wenn es um nicht-bildliche Daten geht, gilt es, eine Argumentationsstruktur aufzubauen, durch die der Zusammenhang zwischen Feature und Label deutlich wird. (Wang, et al., 2019) nennen unter anderem die dedukti-

ve und die induktive Argumentation. Unter deduktiv ist ein „Top-Down“-Vorgehen zu verstehen. Das heißt, dass von einer Grundvoraussetzung ausgehend eine Schlussfolgerung getroffen wird. Bei der induktiven Struktur wird von der Beobachtung aus auf mögliche Ursachen dafür geschlossen. Dies ist ein „Bottom-Up“-Vorgehen.

Die Untersuchung, mit welcher Wahrscheinlichkeit ein Ergebnis von einem Feature abhängig ist, gilt als induktives Vorgehen, da von der Beobachtung auf die Ursache geschlossen wird. Wird einem Arzt beispielsweise die Wahrscheinlichkeit präsentiert, mit der ein Patient eine bestimmte Diagnose erhält, ist es unmöglich die Gründe für die Prognose nachzuvollziehen. Der Arzt muss sich mit den Einflussfaktoren auseinandersetzen, die der Algorithmus als besonders relevant bewertet (Wang, et al., 2019).

Ein weiteres induktives Argumentationsverfahren besteht in der Gruppierung ähnlicher Datenobjekte. Dabei werden Merkmale erarbeitet, in denen sich die Objekte unterscheiden, um allgemeine Konzepte zu entwickeln.

Um die Erklärungen nutzerzentriert darzustellen, bietet es sich an, diese in Form von Fragen und Antworten zum Modell zu gestalten (Wang, et al., 2019). Eine geeignete Frage ist: „Warum hat das Modell so entschieden?“ Der Grund für eine bestimmte Handlung ist für die Nutzer nicht immer auf den ersten Blick erkennbar, sodass eine explizite Nennung hilfreich ist. Außerdem sind Antworten auf folgende weitere Fragen relevant: „Warum hat das Modell nicht etwas anderes entschieden?“ „Warum ist diese Vorgehensweise oder Entscheidung besser in Bezug auf Größen wie Effizienz, Sicherheit und Kosten?“ Diese Fragen sind besonders dann hilfreich, wenn der Algorithmus über Vorgehensweisen, beispielsweise eine organisatorische Planung, entscheidet (Fox, et al., 2017).

Zusammengefasst ergeben sich drei Faktoren, die zu einer nachvollziehbaren Ausgabe beitragen (Wang, et al., 2019): Erstens, die Nennung der relevanten Eingabe-Features zusammen mit der Angabe, ob sie sich positiv oder negativ auf das Ergebnis auswirken. Zweitens können ähnliche

Instanzen aus den Trainingsdaten gruppiert werden, um Muster zu erkennen. Schließlich ist es hilfreich, einzelne Datenobjekte zu präsentieren, indem einige Features mit den Werten genannt werden. Dazu sollte angegeben werden, ob dieser Wert über dem Schwellenwert liegt.

Eine effektive Möglichkeit ist eine Liste mit den zu betrachtenden Input Features und dem Label. Logische Zusammenhänge können in Form von Regeln oder Entscheidungsbäumen formuliert werden. Zur Darstellung von Konzepten und Ontologien bieten sich Graphen an, bei denen Knoten über Kanten miteinander verbunden sind. Dadurch werden das dahinter stehende Konzept und interne Zusammenhänge erläutert.

(Goebel, et al., 2018) stellen in diesem Zusammenhang den „annotated interaction path“ vor. Die relevanten Features werden dabei anhand medizinischer Recherche-Ergebnisse zu einem Wissensgraphen zusammengesetzt. An den Kanten werden dann Annotationen gemacht, die auf medizinische Publikationen mit entsprechenden Belegen verweisen.

Für besonders komplexe Zusammenhänge ist es sinnvoll, diese visuell aufzubereiten (Wang, et al., 2019). Zur Darstellung von Daten bieten sich Diagramme an, die die Transparenz im Prozess fördern. Abbildung 12 ist das Ergebnis einer AI Software, die anhand fiktiver Patientendaten (MIMIC3 Datensatz) eine Diagnose stellt. Die wichtigsten Erklärungen sind hier anschaulich aufbereitet, damit die Diagnose nachvollziehbar ist (Wang, et al., 2019). Unter dem Punkt „System Predicted Risk“ ist ein Beispiel für die Darstellung eines Zusammenhangs in Form eines Balkendiagramms. Es zeigt fünf für den Patienten prognostizierte Risiken, zusammen mit der Wahrscheinlichkeit, dass dieses Risiko eintritt. So ist auf den ersten Blick erkennbar, dass mit großer Wahrscheinlichkeit ein Herzinfarkt („Acute Myocardial Infarction“) auftritt.

Die Diagramme unter dem Punkt „Influence of Vital“ stellen den Einfluss dar, den die zwölf aufgelisteten Vitalwerte auf das Ergebnis haben. Durch eine solche Darstellung der relevantesten Einflussfaktoren kann der Be-

trachter Schlussfolgerungen für weitere mögliche Folgen ziehen. Außerdem ist ersichtlich, wie sich bestimmte Features verhalten müssten, um das Ergebnis in eine andere Richtung zu lenken, weil Features verglichen werden und deren Zusammenhänge nachvollziehbar sind.

Abbildung 12: Erklärungen zur AI Software, die anhand von Patientendaten eine Diagnose stellt. Quelle: (Wang, et al., 2019)

Partial Dependence Plots sind hilfreich, wenn es darum geht, zu analysieren, wie sich die Ausgabe im Zusammenhang mit einem bestimmten Feature verändert. Bei einem solchen Diagramm wird die Eingabe durch einen

Grafen mit der Ausgabe in Verbindung gebracht. Um die Auswirkung eines Features auf die Prognose nachzuvollziehen, wird die Veränderung des Features auf der X-Achse und das Label bzw. die Prognose auf der y-Achse dargestellt. Diese Vorgehensweise ähnelt der in der Mathematik bekannten Darstellungsweise eines Grafen in einem Koordinatensystem. Eine ähnliche Methode ist die Verwendung von Streudiagrammen, welche zur Aufdeckung von Ähnlichkeiten von Objekten dienen.

All die hier betrachteten Methoden helfen den Nutzenden dabei, die Entscheidung eines Algorithmus zu verstehen und nachzuvollziehen. Daraus können Beobachtungen vereinfacht werden, sodass in zukünftigen Situationen Vertrauen in das Modell gesetzt werden kann. Wenn es darum geht, den Benutzer der Software über die eingesetzte KI-Technologie aufzuklären, können Beispiele verdeutlichen, von welchen Faktoren eine automatisierte Entscheidung abhängig ist, wie diese erhoben werden und wie stark deren Einfluss ist (Beining, 2019).

Entwickler zögern oft, Erklärungen über den Algorithmus abzugeben, da Benutzer den Algorithmus durchschauen und manipulieren könnten. Deshalb sollte der Prozess „so konkret wie möglich und so allgemein wie nötig“ erläutert werden (Beining, 2019). Auskünfte über die verwendeten Daten und den Entwicklungsprozess der Software steigern das Vertrauen. Gleichzeitig verdeutlicht die Angabe von Fehlerraten, dass die Ergebnisse überprüft werden müssen (Beining, 2019).

4. Zusammenhang zwischen Fairness und Erklärbarkeit

Fairness und Erklärbarkeit sind zentrale Forschungsthemen in den Bereichen künstliche Intelligenz und Machine Learning. Sie gehören zu den vier ethischen Grundsätzen für vertrauenswürdige Software (Europäische Kommission, 2018) und tragen damit maßgeblich zur Qualität und Akzeptanz der entwickelten Systeme bei. In der Fachliteratur werden beide Grundsätze meistens getrennt voneinander betrachtet, was auch zu getrennten Lösungswegen führt (Sharma, et al., 2019). Wie stehen die Themen in Zusammenhang? Dieses Kapitel fasst die wichtigsten Erkenntnisse aus Kapitel 2 und 3 zusammen, erläutert Gemeinsamkeiten und Unterschiede und analysiert, wie sich die Bereiche ergänzen.

Im gesamten Entwicklungsprozess können Umstände verursacht werden, die zu einer diskriminierenden Prognose führen. Das liegt teilweise am gewählten Modell oder auch am Entwickler selbst. Durch eine vorurteilsbehaftete Auswahl der Features und Labels werden geschützte Merkmale gezwungenermaßen entweder vernachlässigt oder mit einbezogen, obwohl die wahre Entscheidungsgrundlage für die Ausgabe in anderen Features liegt (siehe Kapitel 2.4.3) Dabei spielt das Herausarbeiten der Kausalität zwischen Feature und Label eine große Rolle, um verdeckte Beeinflussungen zu vermeiden. Auch der Kontext, in dem das Modell eingesetzt wird muss gründlich untersucht werden, insbesondere wenn während des Entwicklungsprozesses Änderungen auftreten.

Bei fehlender Transparenz ist das Modell ebenso für den Grad der Nachvollziehbarkeit entscheidend. Neuronale Netze sind zwar sehr präzise, werden jedoch häufig als „Blackbox“-Modelle bezeichnet, weil auf den ersten Blick nicht identifizierbar ist, nach welchen Kriterien eine Entscheidung getroffen wird (Kapitel 3.3). Andererseits sind nachvollziehbare Modelle weniger präzise, sodass bei der Wahl des Modells die Qualitätsanforderungen berücksichtigt werden müssen.

Besonders die Qualität der Daten wirkt sich auf die Fairness des Systems aus. Situationen, in denen Personengruppen bereits benachteiligt und dann in den Trainingsdaten abgebildet werden, führen dazu, dass der Algorithmus diese Vorurteile als Muster erkennt und übernimmt. Weiterhin tragen nicht-repräsentative Daten und qualitative Unterschiede in der Erhebung zur Problematik bei. Durch Proxies können geschützte Merkmale unbemerkt in das Training des Modells einfließen, sodass eine Diskriminierung verdeckt vorhanden ist. Diese Ursachen bezüglich der Trainingsdaten sind im Detail in Kapitel 2.4.1 bis 2.4.4 dargestellt.

Auch für die Förderung der Transparenz sind korrekte Datensätze wichtig, weil daraus im Nachhinein Schlüsse gezogen werden. Die Untersuchung der relevanten Features schafft ein Bild über die Entscheidungsgrundlage des Modells (Kapitel 3.4). Handelt es sich bei diesen Features um Proxies, wird jedoch ein falsches Bild vermittelt, sodass der wahre Einflussfaktor nicht klar ist. Zur Förderung der Transparenz und Aufdeckung von Diskriminierung reicht es nicht, die für die Prognose relevanten Features oberflächlich aufzulisten. Die Analyse der Kausalitäten muss tiefgehend erfolgen, um die Wechselwirkungen innerhalb des Modells zu erörtern.

Besonders die Entwickler wirken während des gesamten Entwicklungsprozesses auf die Qualität des Systems ein. Durch die Auswahl der Features und Label und während der Datenerhebung beeinflussen Vorurteile unbemerkt die Handlungen und Entscheidungen des Entwicklers (Kapitel 2.4.3 und 2.4.4). Gleichzeitig besteht oft ein gewisses Widerstreben, das Modell nachvollziehbar zu gestalten, weil sich daraus aus Sicht des Entwicklers nicht genügend Vorteile ergeben (Kapitel 3.3).

Das zeigt, dass für die Entwicklung fairer und erklärbarer Modelle eine Sensibilisierung der Entwickler stattfinden muss. Sie müssen besonders im Hinblick auf Fairness und Erklärbarkeit über die Ursachen und Lösungsmethoden aufgeklärt werden. Außerdem sollte das Team möglichst vielfältig zusammengesetzt sein, also bestehend aus Menschen mit den

verschiedensten Hintergründen, damit sich die Kenntnisse ergänzen und bereits ein grundlegendes Bewusstsein für die Gefahren vorhanden ist.

Darüber hinaus darf der Algorithmus gemäß den ethischen Grundsätzen nicht autonom handeln (Europäische Kommission, 2018). Machine Learning muss überwacht und kontrolliert stattfinden, wie durch das interaktive Machine Learning nach (Holzinger, 2018). Dabei greift der Entwickler aktiv in den Prozess ein, um den Algorithmus in die richtige Richtung zu lenken.

Sowohl für die Schaffung von Transparenz, als auch für die Verbesserung der Fairness gibt es viele Möglichkeiten für Prävention und Nachbereitung. Ein Beispiel ist die in Kapitel 3.5.2 vorgestellte Vorgehensweise von (Sharma, et al., 2019), bei der Datenobjekte aufgezeigt werden, die möglichst ähnliche Eigenschaften wie das betrachtete Datenobjekt haben, aber ein anderes Label. Dadurch ist ersichtlich, welche Features einen großen Einfluss auf die Prognose haben. Darüber hinaus ergibt sich ein Vorteil für die Förderung der Fairness: Unter Umständen ergibt diese Analyse, dass die betrachtete Person ein anderes Label erhalten hätte, wenn ein geschütztes Merkmal anders wäre. Damit ist der Algorithmus nicht neutral gegenüber allen Personengruppen. Das bedeutet jedoch nicht, dass fehlende Features irrelevant für die Entscheidung sind (Wachter, et al., 2018). Fehlen unter den wichtigsten Features geschützte Merkmale, heißt das also nicht zwingend, dass der Algorithmus fair entscheidet.

Weitere Messungen, die im Rahmen der Erklärbarkeit vorgenommen werden, weisen auf Diskriminierung hin, wenn dabei Ungleichheiten bezüglich verschiedener Personengruppen aufgedeckt werden (Sharma, et al., 2019). Außerdem kann für Modelle, die für die Öffentlichkeit nicht transparent sind, weder die Ursache für Diskriminierung, noch dessen Grad sicher festgestellt werden (Angelino, et al., 2017). Daraus ist zu schließen, dass die Erklärbarkeit von Algorithmen dazu beiträgt, Diskriminierung aufzudecken (Zuiderveen Borgesius, 2018), da genau nachvollziehbar ist, wie die Entscheidung zustande kommt. Damit kann an der richtigen Stelle angesetzt werden, das Modell zu korrigieren und fair zu gestalten.

5. Fazit

Der Begriff Fairness beschreibt die Abwesenheit von Vorurteilen und einer systematischen Benachteiligung von Personengruppen. Im Lauf der Datenerhebung und Entwicklung treten viele Risiken auf, die verzerrte Daten und vorurteilsbehaftete Ausgaben erzeugen.

Eine Ursache für verzerrte Daten besteht dann, wenn in der Realität Vorurteile die Situation beherrschen. Die dadurch vermittelten Grundwahrheiten verfestigen sich im Modell. Zudem sinkt die Qualität des Modells, wenn die Trainingsdaten unausgeglichen sind. Dazu gehört die Unter- oder Überrepräsentation von Personen, Zeiträumen, geografischen Räumen und Ähnlichem. Auch die Auswahl der Features und Labels hat einen maßgeblichen Einfluss auf den Grad der Fairness. Geschützte Merkmale können ungewollt relevant für die Prognose sein, besonders wenn diese verdeckt in Form von Proxy-Variablen auftreten. Unterschiedliche Methoden und Verantwortliche für die Datenerhebung spiegeln sich in qualitativen Unterschieden innerhalb der gesammelten Daten wieder. Ebenso können bei der Aufbereitung der Daten unbemerkt Fehler entstehen, die die Qualität des Modells herabsetzen.

All diese Faktoren verursachen Trainingsdaten, die dem Algorithmus ein falsches und verzerrtes Bild über die Realität vermitteln. Durch eine Bearbeitung der Daten vor dem Training des Modells werden größere Schäden vermieden. Dazu kann der Datensatz durch eine gezielte Änderung der Label wieder ins Gleichgewicht gebracht werden. Da alle Stakeholder eines Projekts bei Entscheidungen und während der Entwicklung durch persönliche Vorurteile geprägt handeln, muss außerdem eine Sensibilisierung der beteiligten Personen stattfinden. Durch die Optimierung und den Einsatz der Software entstehen weitere Risiken für Diskriminierung, sodass auch hier Gegenmaßnahmen eingeleitet werden müssen.

Auch die Transparenz eines Modells ist sowohl von den technischen, als auch den organisatorischen Umständen abhängig. Die Beschaffenheit der

Blackbox-Modelle lässt per se keinen Einblick in die Entscheidungsfindung zu. Es gibt jedoch einige Methoden, die die Qualität und Fairness des Systems überprüfbar machen.

Die Analyse der resultierenden Vorteile ergibt, dass sich durch die Herstellung von Nachvollziehbarkeit Fehler, insbesondere Diskriminierung leichter aufdecken lassen. Deshalb ist die Förderung der Erklärbarkeit eine fundamentale Maßnahme zur Entwicklung vertrauenswürdiger, fairer und präziser Machine Learning Modelle.

5. Abbildungsverzeichnis

6. Tabellenverzeichnis

7. Literaturverzeichnis

ACM U.S. Public Policy Council; ACM Europe Policy Committee, 2017. *Statement on Algorithmic Transparency and Accountability.* [Online]
Available at: http://www.acm.org/binaries/content/assets/public-policy/2017_joint_statement_algorithms.pdf
[Accessed 27 05 2020].

Adadi, A. & Berrada, M., 2018. Peeking inside the black-box: A survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, 17 09, Issue 6, pp. 52138-52160.
https://doi.org/10.1109/ACCESS.2018.2870052

Angelino, E. et al., 2017. Learning certifiably optimal rule lists for categorical data. *The Journal of Machine Learning Research*, 01, 18(1), pp. 8753-8830.
https://dl.acm.org/doi/abs/10.5555/3122009.3290419

Association for Computing Machinery (ACM), 2018. *ACM Code of Ethics and Professional Conduct.* [Online]
Available at: https://www.acm.org/code-of-ethics
[Accessed 27 05 2020].

Baer, T., 2019. *Understand, Manage, and Prevent Algorithmic Bias: A Guide for Business Users and Data Scientists.* Kaufbeuren, Germany: Apress.

Barocas, S. & Selbst, A. D., 2016. Big Data's Disparate Impact. *California Law Review, Vol. 104, Issue No. 3, S. 671-732*, 06.
http://dx.doi.org/10.15779/Z38BG31

Bayati, M. et al., 2014. Data-Driven Decisions for Reducing Readmissions for Heart Failure: General Methodology and Case Study. *PLoS ONE*, 9(10).
https://doi.org/10.1371/journal.pone.0109264

Beining, L., 2019. *Wie Algorithmen verständlich werden - Ideen für Nachvollziehbarkeit von algorithmischen Entscheidungsprozessen für Betroffene.* [Online] Available at: https://www.bertelsmann-stiftung.de/fileadmin/files/BSt/Publikationen/GrauePublikationen/Wie_Algorithmen_verstaendlich_werden_final.pdf
[Accessed 22 06 2020].

Beuth, P., 2017. *Zeit Online.* [Online]
Available at: https://www.zeit.de/digital/internet/2017-07/heiko-maas-algorithmen-regulierung-antidiskriminierungsgesetz
[Accessed 26 05 2020].

Calders, T. & Zliobaite, I., 2013. Why Unbiased Computational Processes Can Lead to Discriminative Procedures. In: *Discrimination and Privacy in the Information Society.* Berlin Heidelberg: Springer Verlag, pp. 43-57. https://doi.org/10.1007/978-3-642-30487-3_3

Cambridge University Press, n.d. *Cambridge Dictionary.* [Online]
Available at: https://dictionary.cambridge.org/dictionary/english-german/bias
[Accessed 23 06 2020].

Die Bundesregierung, 2018. *Strategie Künstliche Intelligenz der Bundesregierung.* [Online]
Available at: https://www.bmbf.de/files/Nationale_KI-Strategie.pdf
[Accessed 27 05 2020].

Europäische Kommission, 2018. *Ethik-Leitlinien für eine vertrauenswürdige KI.* [Online]
Available at: https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60425
[Accessed 27 05 2020].

Fox, M., Long, D. & Magazzeni, D., 2017. Explainable Planning. *IJAC 2017 Workshop for Explainable Artificial Intelligence (XAI)*, 08. https://arxiv.org/pdf/1709.10256

Friedler, S. A. et al., 2019. A comparative study of fairness-enhancing interventions in machine learning. *Proceedings of Conference on Fairness, Accountability, and Transparency*, pp. 329-338. https://doi.org/10.1145/3287560.3287589

Friedman, B. & Nissenbaum, H., 1996. Bias in Computer Systems. *ACM Transactions on Information Systems*, Juli, 14(3), pp. 330-347. https://dl.acm.org/doi/pdf/10.1145/230538.230561

Frochte, J., 2019. *Maschinelles Lernen - Grundlagen und Algorithmen in Python.* 2. aktualisierte Auflage ed. München: Carl Hanser Verlag GmbH & CO. KG.

Fushiki, T. & Maeda, T., 2020. Nonresponse Bias Adjustment in Regression Analysis. *Journal of Statistical Theory and Practice*, 21 02, 14(2), pp. 1-11. https://doi.org/10.1007/s42519-020-0086-z

Gesellschaft für Informatik, 2018. *Technische und rechtliche Betrachtungen algorithmischer Entscheidungsverfahren,* Berlin: Sachverständigenrat für Verbraucherfragen. https://www.svr-verbraucherfragen.de/wp-content/uploads/GI_Studie_Algorithmenregulierung.pdf [Accessed 06.07.2020]

Goebel, R. et al., 2018. Explainable AI: The New 42?. *Machine Learning and Knowledge Extraction*, pp. 295-303. https://doi.org/10.1007/978-3-319-99740-7_21

Gunning, D. & Aha, D. W., 2019. DARPA's Explainable Artificial Intelligence Program. *AI Magazine*, 40(2), pp. 44-58. https://doi.org/10.1145/3301275.3308446

Habibi Aghdam, H. & Jahani Heravi, E., 2017. *Guide to convolutional neural networks.* s.l.:Springer Verlag. https://doi.org/10.1007/978-3-319-57550-6

Holzinger, A., 2018. Explainable AI (ex-AI). *Informatik Spektrum*, 03 04, Issue 41, pp. 138-143. https://doi.org/10.1007/s00287-018-1102-5

Holzinger, A., 2018. From Machine Learning to Explainable AI. *2018 World Symposium on Digital Intelligence for Systems and Machines (DISA)*, pp. 55-66. https://doi.org/10.1109/DISA.2018.8490530

International Organization for Standardization, n.d. *ISO/IEC AWI TR 24027.* [Online] Available at: https://www.iso.org/standard/77607.html [Accessed 27 05 2020].

Kamiran, F., Calders, T. & Pechenizkiy, M., 2013. Techniques for Discrimination-Free Predictive Models. In: *Discrimination & Privacy in the Information Society.* Berlin Heidelberg: Springer-Verlag, pp. 223-239.

https://doi.org/10.1007/978-3-642-30487-3_12

Kolodiazhnyi, K., 2020. *Hands-On Machine Learning with C++.* Birmingham, Mumbai: Packt Publishing.

London, A. J., 2019. Artificial Intelligence and Black-Box Medical Decisions: Accuracy versus Explainability. *Hastings Center Report*, 21 02, 49(1), pp. 15-21. https://doi.org/10.1002/hast.973

Munoz, A., 2014. *Machine Learning and Optimization.* [Online]
Available at:
https://pdfs.semanticscholar.org/7fbb/a79630b5a09dd66ab13f00c3aefaa56cf268.pdf

Orwat, C., 2019. *Diskriminierungsrisiken durch Verwendung von Algorithmen.* [Online]
Available at:
https://www.antidiskriminierungsstelle.de/SharedDocs/Downloads/DE/publikationen/Expertisen/Studie_Diskriminierungsrisiken_durch_Verwendung_von_Algorithmen.pdf?__blob=publicationFile&v=5
[Accessed 01 06 2020].

Petkovic, D., Altmann, R., Wong, M. & Vigil, A., 2018. Improving the explainability of Random Forest classifier – user centered approach. *Pacific Symposium on Biocomputing*, pp. 204-215.
https://doi.org/10.1142/9789813235533_0019

Ren, X. et al., 2019. Neural Network based detection of self-admitted technical debt: From performance to explainability. *ACM Transactions on Software Engineering and Methodology*, 07, 28(3), pp. 1-45.
https://doi.org/10.1145/3324916

Samek, W., Wiegand, T. & Müller, K.-R., 2017. Explainable artificial intelligence: understanding, visualizing and interpreting deep learning models. *arXiv preprint*.
https://arxiv.org/abs/1708.08296v1

Sharma, S., Henderson, J. & Ghosh, J., 2019. CERTIFAI: Counterfactual Explanations for Robustness, Transparency, Interpretability, and Fairness of Artificial Intelligence models. 20 05.

https://arxiv.org/abs/1905.07857v1

Shield, C. & Abbany, Z., 2020. *Epidemiologe Kurth: Corona-Zahlen sind nicht vergleichbar.* [Online]
Available at: https://p.dw.com/p/3aGV4
[Accessed 25 06 2020].

Simmank, J., 2018. *Das Auto, das entscheiden muss, ob es Alte oder Kinder überfährt.* [Online]
Available at: https://www.zeit.de/digital/2018-10/autonomes-fahren-kuenstliche-intelligenz-moralisches-dilemma-unfall
[Accessed 23 11 2020].

Suresh, H. & Guttag, J. V., 2020. *A Framework for Understanding Unintended Consequences of Machine Learning.*
https://arxiv.org/abs/1901.10002

Tickle, A. B., Andrews, R., Golea, M. & Diederich, J., 1998. The truth is in there: directions and challenges in extracting rules from trained artificial neural networks. *IEEE Transactions on Neural Networks*, 11, 9(6), pp. 1057-1068.
https://doi.org/10.1109/72.728352

Vincent, J., 2018. *Amazon reportedly scraps internal AI recruiting tool that was biased against women.* [Online]
Available at: https://www.theverge.com/2018/10/10/17958784/ai-recruiting-tool-bias-amazon-report
[Accessed 19 11 2020].

Wachter, S., Mittelstadt, B. & Russell, C., 2018. Counterfactual Explanations without opening the black box: Automated Decisions and the GDPR. *Harvard Journal of Law & Technology*, Issue 31, p. 841 ff..
https://arxiv.org/ftp/arxiv/papers/1711/1711.00399.pdf

Wang, D., Yang, Q., Abdul, A. & Lim, B. Y., 2019. Designing Theory-Driven User-Centric Explainable AI. *CHI '19: Proceedings of the 2019 CHI Conference on Human Factors in Computing*, 05, pp. 1-15.
https://doi.org/10.1145/3290605.3300831

Wang, T. et al., 2019. Balanced Datasets Are Not Enough: Estimating and Mitigating Gender Bias in Deep Image Representations. *Proceedings of the IEEE International Conference on Computer Vision*, pp. 5310-5319. https://doi.org/10.1109/ICCV.2019.00541

Wehner, M. & Köchling, A., 2020. Learning Analytics und Dsirkiminierung. *DELFI 2020 - Die 18. Fachtagung Bildungstechnologien der Gesellschaft für Informatik e.V.*, pp. 369-370. https://dl.gi.de/handle/20.500.12116/34192 [Accessed 26.11.2020]

Windzio, M., 2013. *Regressionsmodelle für Zustände und Ereignisse.* Wiesbaden: Springer Verlag. https://doi.org/10.1007/978-3-531-18852-2_2

Zielke, T., 2020. *Is Artificial Intelligence Ready for Standardization?.* [Online] Available at: https://www.researchgate.net/publication/341616218_Is_Artificial_Intelligence_Ready_for_Standardization [Accessed 10 08 2020].

Zuiderveen Borgesius, F., 2018. *Discrimination, artificial intelligence, and algorithmic decision-making.* [Online] Available at: https://rm.coe.int/discrimination-artificial-intelligence-and-algorithmic-decision-making/1680925d73 [Accessed 10 08 2020].