HOCHSCHULE DÜSSELDORF
-FACHBEREICH SOZIAL- UND KULTURWISSENSCHAFTEN-

MASTERTHESIS ZUM THEMA

DATEN ALS HEBEL FÜR FEMINISTISCHE KI-ANWENDUNGEN

Erstgutachter: Prof. Dr. Christian Voigt
Zweitgutachter: Carsten Heisterkamp
vorgelegt von: Özden Şenarslan
vorgelegt am: 20. Juli 2022

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Anhangstabellenverzeichnis

Hinweis: Die beigefügte CD zu dieser Thesis enthält die elektronische Version der vorliegenden Arbeit und das Datenmaterial für die systematische Literaturanalyse. Das Datenmaterial besteht aus einer Word Datei mit den 321 codierten Segmenten sowie die dazugehörigen 28 Dokumente in PDF.

Abkürzungsverzeichnis

ABM	Agent Based Models
ADM Systems	Algorithmic Decision Making Systems
AI	Artificial Intelligence
AIC4	Artificial Intelligence Cloud Service Compliance Criteria Catalogue
AIES	Artificial Intelligence, Ethics and Society
ANN	Artificial Neural Networks
BSI	Bundesamt für Sicherheit in der Informationstechnik
DAO	Distributed Autonomous Organizations
DECODE	Decentralised Citizen-owned Data Ecosystems
DisCO	Distributed Cooperative Organization
DL	Deep Learning
DSGVO	Datenschutz Grundverordnung
EU	Europäische Union
HGB	Handelsgesetzbuch
IoT	Internet of Things
KI	Künstliche Intelligenz
KITE	KI Thinktank Female Entrepreneurship
KNN	Künstliche Neuronale Netze
ML	Maschinelles Lernen
PICO-Schema	Population, Intervention, Comparison und Outcome - Schema
SLR	Systematic Literature Reviews
TL	Tiefes Lernen

1. Einleitung

Anwendungen der Künstlichen Intelligenz (KI) mit Maschinelles Lernen (ML) sind aus dem Alltag nicht mehr wegzudenken und ein viel diskutiertes Phänomen, das für die Bewältigung der exponentiell wachsenden digitalen Daten zur Informationsverarbeitung eingesetzt wird. Digitale Daten von Personen sind von wertschöpfendem Interesse für verschiedene Akteur*innen wie Regierungen, Institutionen, Organisationen oder Unternehmen und spielen eine zentrale und aktive Rolle in KI-Anwendungen.

1.1 Relevanz und Motivation

Laut der Datenethikkommission (Datenethikkommission 2019, S. 44) soll die Datenhoheit von personenbezogenen Daten bei denen liegen, die sie erzeugt haben. Die Verfügungsbefugnis über Daten wäre über die digitale Selbstbestimmung geregelt. Diese sei „die Kompetenz, selbst zu bestimmen, mit welchen Inhalten jemand in Beziehung zu seiner Umwelt tritt und wie jemand die eigene Persönlichkeit interaktiv entfaltet“. Solche digitale Selbstbestimmung umfasst nach der Datenethikkommission auch die Datenmonetarisierung von eigenen Datenbeständen sowie Daten von eigenen Internet of Things (deutsch: Internet der Dinge; Abk.: IoT) Geräten. Somit lässt sich mit dieser Formulierung die digitale Selbstbestimmung auch für Unternehmen als juristische Personen miteinbeziehen, die ein hochgradig ökonomisches Interesse an personenbezogenen Daten haben, um ihre Produkte oder Dienstleistungen an die Konsumierenden zu verkaufen. Daher sollen, so die Datenethikkommission, Konsumierenden bewusste Konsumentscheidungen ermöglicht werden, welche die Voraussetzungen zur „optimalen Ressourcenallokation und Wohlfahrtsmaximierung innerhalb der Volkswirtschaft“ sei.

Um die „Volkswirtschaft zu aktivieren“, hat die Bundesregierung (2021, 17 ff.) ihre Datenstrategie veröffentlicht. Hiernach werden Gestaltungsmöglichkeiten gesucht, wie einheitliche und widerspruchsfreie Anonymisierung und technischer Datenschutz bewerkstelligt werden kann, um Innovationen voranzutreiben. Aus diesem Grund wird geprüft, „ob und wie die Datenschutzaufsicht für den nicht-öffentlichen Bereich verbessert werden kann“. Ziel ist es, bereits bei der Datenerhebung rechtliche und technische Maßnahmen zu gestalten, die die „Depersonalisierung von Daten“ ermöglichen. Unter vielen Gestaltungsmöglichkeiten sowie Fördermaßnahmen enthält die Datenstrategie der Bundesregierung die Bestrebung, Datentreuhandmodelle zu etablieren. Treuhändisch Datenbeauftragte sollen neutral und wirtschaftlich unabhängig in der Datensphäre operieren. Ihre Aufgabe soll die Ermöglichung des zügigen „Datenteilens und -nutzens“ sein.

Aus dem Vorhaben mittels Datenökonomie „nachhaltiges Wachstum und Wohlstand durch Datennutzung [zu] fördern“ (Die Bundesregierung 2021, S. 6) und der dahinter stehenden Datenstrategie der Bundesregierung lässt sich schlussfolgern, dass Daten einen immens hohen Wert haben können. Dies ist auch ein Indikator dafür, welche Macht KI-Anwendungen aufgrund von Daten haben können, wenn sie bspw. zur Entscheidung von Teilhabemöglichkeit zu Ressourcenallokation eingesetzt werden.

Weitere Bestrebungen sind der AIC4 (Artificial Intelligence Cloud Service Compliance Criteria Catalogue). AIC4 ist ein Kriterienkatalog, der die Mindestanforderungen an die sichere Verwendung von Methoden des ML in Cloud-Diensten mit KI-Anwendungen spezifizieren soll (BSI 2021b, o.S.). Der Kriterienkatalog bietet für Unternehmen die Möglichkeit sich in einem unabhängigen Audit auf Basis standardisierter Prüfung, einer Risikoanalyse zu unterziehen, ob die eingesetzte KI-Anwendung sicher ist (BSI 2021a, S. 12). Die formulierten Kriterien sind nicht verpflichtend und haben auch keine Konsequenz bei Nichteinhaltung (BSI 2021b, o.S.). Dieser Kriterienkatalog für KI-Anwendungen ist in dieser Form einmalig:

> “Up to now there are no other international or national standards available proving to be equally suitable for conducting such an audit” (BSI 2021a, S. 11).

Die Risikoanalyse ist in folgende acht Bereiche unterteilt: Vorläufige Kriterien, Sicherheit & Robustheit, Leistung & Funktionalität, Zuverlässigkeit, Datenqualität, Datenmanagement und Bias (ebd., S. 8).

Der letzte Punkt sei, so das BSI, oft mit moralischen oder ethischen Fragen verbunden, wie der fairen Behandlung von Individuen oder sozialen Gruppen. Es positioniert sich jedoch nicht inhaltlich diesen Fragen (BSI 2021a, S. 42). Während das Forschungsgebiet KI weltweit an Bedeutung zunimmt, die in vielen Forschungs- und Entwicklungsaktivitäten münden, sind ethische Fragen zu bedenklichen KI-Anwendungen noch im Anfangsstadium (► vgl. Kap. 3.3.2).

Laut des AI Index Reports (Zhang et al. 2022, 17 ff.) der bereits zum vierten Mal in Folge erscheint, ist die Anzahl der KI-Veröffentlichungen in den letzten 20 Jahren drastisch gestiegen. Allein zwischen 2019 bis 2020 stieg die Anzahl der KI-Veröffentlichungen um 34,5 Prozent, während zwischen 2018 bis 2019 der Anstieg noch 19,6 Prozent betrug. Der größte Anteil der KI-Publikationen, die von Expert*innen begutachtet wurden (engl. peer-reviewed) stammen von akademischen Einrichtungen aus China, den Vereinigten Staaten (USA) und der Europäischen Union (EU). Diese Großmächte sind auch diejenigen, die die höchsten Investitionen in KI-Forschung getätigt haben. Dem Bericht zufolge hat das Aufkommen von KI-Konferenzen und

Preprint-Archiven[1] die Verbreitung von Forschung und wissenschaftlicher Kommunikation erweitert, was zur Folge hat, dass die Erkenntnisse in der KI-Forschungsdisziplin sehr vielfältig und schnelllebig sind. Umso überraschender ist dem Bericht zu Folge, dass wenig Forschungsinformationen zur Bewältigung bestehender ethischer Herausforderungen von unvoreingenommen und diskriminierungsfreien KI-Anwendungen und die dazugehörigen Entwicklungen gibt. Grund hierfür sei, dass im Bereich der KI-Ethik ein allgemeiner Maßstab für den Vergleich von Messwerten (engl. Benchmark) fehlt, um die Beziehung zwischen zivilgesellschaftlichen Diskussionen über KI-Anwendungen und die dazugehörigen technischen Entwicklungen zu messen oder zu bewerten. Daher sei mit der zunehmenden Verbreitung von Innovationen, die auf KI beruhen, die ethischen Herausforderungen von KI-Anwendungen immer offensichtlicher und würden auf dem Prüfstand stehen (Zhang et al. 2022, 17 ff.).

Die Motivation der vorliegenden Arbeit ist daher, mit systematischer Literaturrecherche und anschließender qualitativer Inhaltsanalyse in der schnelllebigen KI-Fachdisziplin herauszufinden, was über diskriminierungsfreie und unvoreingenommene Daten in datengetriebenen KI-Anwendungen mit ML bekannt ist, da wie bereits oben aufgezeigt, ethische Fragen immer noch offen sind, wenn es um den Schutz von personenbezogenen Daten geht. All die vorgestellten Veröffentlichung, wie die der Datenstrategie der Bundesregierung, die Vorschläge der Datenethikkommission und der Kriterienkatalog AIC4 des BSI sowie der AI Index Report zeigen auf, dass für personenbezogenen Daten in KI-Anwendungen keine hinreichende Lösung gefunden ist, oder noch in der Findungsphase sind. Die Relevanz der vorliegenden Arbeit lässt sich aus diesen Veröffentlichungen ableiten. Daten werden als „Wirtschaftsgut angesehen, das Wertschöpfung ermöglicht“ (Die Bundesregierung 2021, S. 4), jedoch stehen das ökonomische Interesse an personenbezogenen Daten und die demokratischen Werte der westlichen Länder im Widerspruch.

1.2 Ziel der systematischen Literaturrecherche

Das Ziel der vorliegenden Arbeit besteht darin, bereits vorhandene Publikationen zu diskriminierungsfreien und unvoreingenommenen Daten, die in KI-Anwendungen zum Einsatz kommen, mithilfe der systematischen Literaturanalyse zu identifizieren und zusammenzufassen, um erstens Schlüsse für feministisch veranlagte datengetriebene KI-Anwendungen zu ziehen, zweitens einen Überblick über den aktuellen Stand der Forschung zu geben und drittens mögliche Desiderate aufzudecken. Hierzu wird mithilfe

[1] Preprints sind vorläufige Versionen wissenschaftlicher Manuskripte, die Forscher auf Online-Plattformen, sogenannten Preprint-Servern, veröffentlichen, bevor sie von Fachkolleg*innen begutachtet und in einer wissenschaftlichen Zeitschrift veröffentlicht werden. Preprint-Server sind öffentlich zugängliche Online-Archive, die Preprints und die dazugehörigen Daten aufbewahren.

der systematischen Literatursuche Publikationen datenbankgestützt und Algorithmus geleitet recherchiert, methodisch gesammelt, nach festgelegten Einschluss- und Ausschlusskriterien bewertet und aussortiert. Die daraus resultierende Datenextraktion wird mithilfe von qualitativen Auswertungsverfahren analysiert und dargestellt. Da es sich, wie Eingangs erörtert, um einen Forschungsbereich handelt, der im Anfangsstadium ist, lässt sich noch kein allgemeingültiger Ansatz für diskriminierungsfreie und unvoreingenommene Daten aufstellen. Deshalb liegt die Betrachtung der vorliegenden Arbeit schwerpunktmäßig auf den vorhandenen Ansätzen, die den ethischen Aspekt von diskriminierungsfreien und unvoreingenommenen Daten in feministisch veranlagten KI-Anwendungen mit ML thematisieren, wobei auch die Ursachen von diskriminierenden und voreingenommenen Daten untersucht werden, um mögliche Lösungsansätze zu verstehen.

1.3 Aufbau

Die vorliegende Arbeit gliedert sich in sechs Kapitel. Kapitel zwei führt in das Themenkomplex KI mit den dazugehörigen zentralen Begriffen ein und zeigt den aktuellen Forschungsstand auf. Im Kapitel drei wird der theoretische Bezugsrahmen festgesetzt, der die Rolle und die Bedeutung der Daten in KI-Anwendungen thematisiert. Darauffolgend wird die Entstehungsgeschichte des Feminismus dargestellt und die Notwendigkeit aufgezeigt, dass die Theorie des Schwarzen Feminismus für eine feministische KI-Anwendung am geeignetsten ist. Zum Abschluss dieses Kapitels werden die ethischen Werte einer KI-Anwendung diskutiert und eine Verknüpfung zwischen den Themen KI-Anwendung, Daten und Feminismus hergestellt und schlussendlich ein Framework für die Interpretation der erhobenen Daten gesetzt. Mit der Methodik der systematischen Literaturanalyse befasst sich Kapitel vier. In diesem Kapitel wird die Forschungsfrage definiert sowie der gesamte Forschungsablauf protokolliert und erläutert. Das Protokoll beinhaltet die Identifizierung der Datenbanken, die sich für die Literaturanalyse eignen, die Festlegung der Suchstrategien sowie die Definition der Ein- und Ausschlusskriterien. Ferner werden die erhobenen Publikationen anhand eines selbstentwickelten Kriterienkatalogs bewertet und mit den festgelegten Ausschlusskriterien aussortiertet. Mit der Software MAXQDA werden die erhobenen Publikationen einer qualitativen Textanalyse unterzogen und kodiert. Die Forschungsergebnisse werden anschließend in Kapitel fünf dargestellt, bezugnehmend zur dargestellten Theorie interpretiert und diskutiert. Im Kapitel sechs wird die vorliegende Arbeit mit einem Fazit abgeschlossen. Es umfasst die Beantwortung der Forschungsfragen und die dazugehörige Zusammenfassung, zudem wird eine Reflexion über die Untersuchung durchgeführt und mit einem Ausblick wird die vorliegende Arbeit beendet.

2. Zentrale Begriffe und Forschungstand

In den folgenden Kapiteln werden zentrale Begriffe entlang ihres Entstehungskontexts mit Bezug auf den Forschungsstand hergestellt und diskutiert.

2.1 Künstliche Intelligenz

Die Geburtsstunde der KI begann mit Alain Turings Frage „Können Maschinen denken?“ (Turing 1950, S. 1). Mit der darauffolgenden Konferenz in Dartmouth wurde nicht nur das akademische Fachgebiet für ‚Artificial Intelligence‘ (engl. Abk. AI, deutsch KI) gegründet, sondern auch geprägt (McCarthy et al. 1955, S. 2). Seit diesen Anfängen erfährt das Forschungsfeld der KI immer wieder Phasen der Hypes (Forni et al. 2017, o.S.), die geprägt sind von optimistischen Vorhersagen der Forschenden und Desillusionen der Fördernden (Lighthill 1972, o.S.). Infolgedessen wurden die Phasen KI-Winter genannt, wenn die Forschung zum Erliegen kam, weil die Grenzen der vorhandenen Technologien aufgrund von Rechenleistungen das Vorhaben der Forschenden ausbremsten oder durch leistungsfähige Technologien einen Schub erfuhren. Ideen, die zuvor nicht möglich waren, wurden durch technologische Digitalisierungsfortschritte und insbesondere mit dem Aufkommen des Internets umgesetzt, was mit Innovationen einherging, womit der KI-Frühling eingeläutet wurde. In allen Phasen der KI-Entwicklung sind kontroverse Diskussionen von starken Emotionen bestimmt und beinhalten eine anthropomorphisierende Semantik sowie verschiedenartige Definitionen von *Künstliche Intelligenz*. Ursächlich hierfür ist der Begriff *Künstliche Intelligenz*. Es kursieren viele Definitionen von KI (Richter et al. 2019, S. 35). Je nachdem aus welcher Wissenschaftsperspektive oder Hype-Phasen der KI betrachtet wird, variieren auch die Definitionen (ebd., S. 36). Ein weitere Schwierigkeit bei der Definition der KI besteht darin, dass für die menschliche Intelligenz keine „explizit-verbale Definition“ gibt, obwohl diese, das „am besten erforschte Merkmal der Psychologie“ ist (Rost 2013, S. 12). Intelligenz ist die Fähigkeit …

> „.. to do the right thing at the right time, in a context where doing nothing (making no change in behavior) would be worse (…) Intelligence is the subset of computation that transforms a context into action.” (Bryson 2019, S. 3).

Die menschliche Intelligenz, die als Vorbild für künstliche Intelligenz dient:

> „… is a term used to describe (typically digital) artifacts that extend any of the capacities related to natural intelligence” (ebd.).

Für die vorliegende Arbeit wird die Definition ausgewählt, weil sie darauf hinweist, dass auch eine KI als Teilgebiet der Informatik ein Rechenhilfsmittel ist, die auf der Grundlage von Berechnungen Informationen verarbeitet, wobei die menschliche Informationsverarbeitung als Vorbild dient. Auch die kognitive Leistung der

menschlichen Intelligenz beinhaltet „all jene Prozesse, durch die (...), [die Informationen] umgesetzt, reduziert, weiterverarbeitet, gespeichert, wieder hervorgeholt und schließlich benutzt wird“ (Neisser 1974, S. 19), die in ähnlicher Manier auch auf dem Artefakt Computer stattfinden, worauf KI-Anwendungen mittels Programmiersprachen realisiert werden. Hierzu benötigen KI-Anwendungen Unmengen an Daten zur Informationsverarbeitung, was im weiteren Verlauf der Thesis aufgezeigt wird.

Zusammenfassend lässt sich feststellen, dass die KI-Forschung von Höhen und Tiefen geprägt war, aber letztendlich immer wieder in verschiedenen Epochen der Digitalisierung durch die Entwicklung der leistungsfähigen Technologien zum Vorschein kam, die für die Datenproduktion und -nutzung den Weg geebnet haben. Es lässt sich nicht von der Hand weisen, dass die aktuelle KI-Forschung mitten im KI-Frühling ist. Dies wird bspw. durch die privaten Investitionen im Jahr 2021 deutlich, die sich auf insgesamt 93,5 Milliarden Dollar beliefen; im Jahr zuvor betrugen diese noch die Hälfte (Zhang et al. 2022, S. 3) und 2019 waren bereits über 94 % der KI-Patente in Ländern mit hohem Einkommen angemeldet, wobei die ersten fünf Plätze die Länder USA, Japan, Frankreich, Kanada und Deutschland belegen (Perrault et al. 2019, S. 32). Folglich wird durch die Platzierung der Länder deutlich, wer die Beforschung der KI vorantreibt und auch bestimmt, in welche Richtung die Entwicklung der KI geht. Der eindeutige Fokus auf die ökonomischen Vorteile lässt die ethischen Aspekte der KI-Anwendungen weitestgehend als Nischendasein erscheinen. Werden die medialen Diskurse betrachtet, so attestiert eine Studie der Bertelsmann Stiftung (2020) mangelnde Vielfalt, da in den Berichterstattungen die wirtschaftlichen Interessen dominieren, wohingegen die Interessen der Zivilgesellschaft oder politisch Agierenden kaum vertreten sind (Fischer et al., S. 6, 2021). Demnach sind die „gesellschaftliche[n] Chancen für das Gemeinwohl und öffentlich finanzierte Einsatzgebiete wie Medizin oder Bildung (…) seltener thematisiert“ (ebd., S. 7). Hierbei wird die Abwesenheit derjenigen deutlich, um die es geht. Die, die Daten produzieren werden unter dem Aspekt des Wohlbefindens nicht priorisiert, von denen die Ökonomie profitiert. Was genau letztendlich KI ist, bleibt diffus. Während sozialwissenschaftliche bzw. ethische Diskussionen den Begriff KI verwenden, ist im Ingenieur*innenwesen die Rede vom ML. Der historische Verlauf des Forschungsgebiets zeigt auch auf, dass der Begriff *Künstliche Intelligenz* oftmals allgemein gehalten wurde, um die wissenschaftlichen Bestrebungen zu vermarkten (Bauberger 2020, S. 26), weshalb auch verschiedenartige KI die Diskurse dominieren. Deshalb wird im nächsten Kapitel auf die Arten der KI eingegangen.

2.2 Die Arten der KI – schwach, stark & superintelligent

Der Wunsch der Menschheit eine superintelligente Maschine zu erschaffen, die die Fähigkeiten des Menschen imitiert und genauso intelligent ist, wenn nicht intelligenter (Mayor 2020, S. 7), soll laut dem Director of Engineering bei Google im Jahr 2045 eintreten (Kurzweil 2013, 56 f.). Er argumentiert diese Prophezeiung mit dem Moor'schen Gesetz. Moore, der Mitbegründer der Firma Intel, sagte im Jahr 1965 voraus, dass die Anzahl der Transistoren auf einem Mikrochip sich jedes Jahr verdoppelt, während die Produktionskosten sinken (Moore 1965, S. 2). 1965 fanden noch 50 Transistoren auf einem Mikrochip ihren Platz (ebd.), inzwischen haben 50 Milliarden Transistoren Platz auf einem fingernagelgroßen Chip (IBM 2021, o.S.). Damit eröffnen sich laut IBM mehr Innovationsmöglichkeiten für KI-Technologien (ebd.). Mit dem ständigen Fortschritt wäre laut Walsh (2018, S. 172) die Idee der technologischen Singularität nicht ganz abwegig. Bereits 1993 prophezeit Vinge (1993, S. 1) für die kommenden 30 Jahre würde die technologische Singularität eintreten. Ihm nach ist die menschliche Ära kurze Zeit später zu Ende, wenn die technischen Mittel zur Verfügung stehen, um eine übermenschliche Intelligenz zu schaffen (ebd.). Aktuell entspricht der technologische Fortschritt weder der menschlichen Intelligenz, was einer starken KI entspräche (engl. strong AI), noch einer übermenschlichen KI, also der einer superintelligenten KI (engl. general AI) (Richter et al. 2019, S. 37). Die heutigen KI-Anwendungen, die in vielen Lebensbereichen der Gesellschaft ihren Einsatz findet, entsprechen der einer schwachen KI (engl. weak/narrow AI) (ebd.). Programmiert wird diese mit den Computersprachen wie Python, C++, Java, Lisp oder Prolog, mit denen Algorithmen codiert werden. Besonders gut ist die schwache KI beim Prozessieren von riesigen Datenmengen, wobei ihr Kernmerkmal das Lösen von spezialisierten Aufgaben ist, für die sie mit den dazugehörigen Algorithmen programmiert wird (ebd.). Defacto kann eine schwache KI, die auf eine bestimmte Aufgabe respektive einen Anwendungsbereich trainiert wurde, nicht die Problemstellungen anderer Anwendungsbereiche lösen. Bspw. kann eine KI-Anwendung, die im medizinischen Bereich eingesetzt wird, um Diagnosen über bestimmte Krankheiten abzugeben, nicht ohne Weiteres zum Erschaffen von Kompositionen der klassischen Musik eingesetzt werden. Walsh (2018, S. 17) nennt die Software der schwachen KI ‚dumme Gelehrte',[2] die über eine ‚Inselbegabung' verfügt. Intelligent wird die schwache KI als Software deshalb bezeichnet, weil sie in der Lage ist, menschliche kognitive Einzelfähigkeit als *Inselbegabte* zu ersetzen, die bisher nur den Menschen zugesprochen wurden (Zweig 2019b, S. 318). Diese Einzelfähigkeiten werden nicht statisch programmiert, sondern

[2] Original in engl. ‚Idiot savant'.

aus Daten durch ML in KI-Anwendungen erlernt, deren Funktionsweise und die algorithmische Relevanz von KI-Anwendungen im Folgenden skizziert wird.

2.3 Funktionsweise und algorithmische Relevanz von KI-Anwendungen

Wird die Historie der Menschheit betrachtet, so ist festzustellen, dass entlang ihrer Entwicklung Menschen immer schon danach bestrebt waren, soziale Prozesse durch Hilfsmittel zu verbessern, die sie in ihrer Tätigkeit unterstützt. Der Abakus zählt zu den ältesten Rechenhilfsmitteln (Beauclair 1968, S. 11) und mündet in der Fortschreibung der Künstlichen Intelligenz (KI). Auch die KI als Teilgebiet der Informatik ist ein Rechenhilfsmittel, welches mithilfe von mathematischen Funktionen die heutigen Probleme löst. KI-Anwendungen sind „*von Menschen entwickelte Softwaresysteme (…), die in Bezug auf ein komplexes Ziel auf (…) digitaler Ebene handeln“* (HEG-KI 2018, S. 6). Mithilfe des Artefakts Computer wird, das Handeln für KI-Anwendungen, auf digitaler Ebene ermöglicht (Bryson 2019, S. 3). Für KI-Anwendungen bedeutet *handeln,* dass sie entscheidungsunterstützend oder -treffend eingesetzt werden. KI-Anwendungen, die unterstützend eingesetzt werden (Decision Support Systems), sollen Expert*innen bspw. im Gesundheitsbereich den Ärzten erlauben, in ihren Diagnosen auf weiteres Wissen zurückzugreifen (Zweig 2018, S. 12). Gemeinsam mit Expert*innen und Programmierenden werden die Lösungsstrategien durch einen Algorithmus kodiert, der eine Reihe von auszuführenden Handlungsanweisungen beinhaltet, „um die Eingabe in eine Ausgabe zu transformieren“ (Alpaydın 2022, S. 1). Ihre Begrenztheit liegt jedoch darin, dass nicht alles Wissen der Welt und die dazugehörigen Problemlösungsstrategien in Algorithmen modelliert und operationalisiert werden können. Was es jedoch gibt, sind Unmengen an Daten (engl. Big Data), die manuell gar nicht informationstechnisch verarbeitet werden können sowie die dazugehörigen Probleme oder Aufgaben. Hierzu werden ML mit lernenden Algorithmen eingesetzt.

Im Kontext von KI-Anwendungen wird eine formale Definition für einen lernenden Algorithmus als mathematisches Konstrukt vorgeschlagen (Mittelstadt et al. 2016, S. 2). Sie besagt, dass

> „An algorithm is a finite, abstract, effective, compound control structure, imperatively given, accomplishing a given purpose under given provisions” (Hill 2016, S. 47).

Solch eine allgemeine Definition beschreibt die Vorgehensweise von ML-System mit lernenden Algorithmen, die die Problemstellungen durch Lernen lösen, ohne, dass der Lösungsweg durch Programmierende vorgegeben wird. Dieser Ansatz der lernenden Algorithmen ist, der des algorithmischen Entscheidungssystems (Algorithmic decision making systems, ADM Systeme). ADM Systeme verwenden zu Beginn einen

sehr allgemeinen Algorithmus mit vielen freien Parametern, der durch Programmierende im Designprozess ausgewählt und modifiziert wird (Alpaydın 2022, S. 1). Dieser Algorithmus lernt aus Trainingsdaten, „wie Personen in der Vergangenheit kategorisiert wurden [Klassifikationsmethode] oder welches Verhalten sie zeigten [Scoring- oder Regressionsmethode]“ (Zweig 2018, S. 13). Lernen bedeutet in diesem Fall, dass der Algorithmus unter Verwendung der Trainingsdaten ausgeführt und entsprechend durch die Modifizierung der Parameter optimiert wird (Alpaydın 2022, S. 3). Aus den gelernten Trainingsdaten wird ein zweiter Algorithmus mit Entscheidungsregeln abstrahiert und abgespeichert (Krafft et al. 2019, S. 8), welcher der eigentliche Algorithmus (engl. Gradient Boosting-Algorithmus) ist, der dann die Entscheidungen trifft. Die Art und Weise des zweiten Algorithmus ist „meist extrem simpel“ aufgebaut (Zweig 2018, S. 13) und ist der eigentliche Algorithmus, der die konkreten Aufgaben löst (Alpaydın 2022, S. 1). Auf den Punkt gebracht, geht es bei ML darum, die richtigen Trainingsdaten zu verwenden, um die richtigen Algorithmen zu erstellen, die die richtigen Aufgaben erfüllen (Flach 2012, S. 12). Ungerechtfertigte Entscheidungen, die durch KI-Anwendungen getroffen werden können, hängen folglich nicht unbedingt von der Software ab, weil

> „die zugrunde liegenden Algorithmen, mit denen ein Modell [KI-Anwendung] trainiert wird, relativ einfach zu spezifizieren sind“ und „[i]n vielen Fällen sind sie bereits seit Jahren in Softwarepackages professionell implementiert und oft ist auch der [Programm]Code öffentlich zugänglich“ (Zweig 2018, S. 22).

Damit ist gemeint, dass Algorithmen, die in KI-Anwendungen verwendet werden, nicht mehr unbedingt selbst programmiert werden müssen[3] und erprobt oder ggf. ausgebessert worden sind (Otte 2019, S. 447).

> Laut Zweig (2018, S. 15) „können Algorithmen [in KI-Anwendungen] diskriminierungsfrei entscheiden“, wenn sie „keinen Input als Trainingsdatensatz bekommen, der eine Diskriminierung beinhaltet, und (…) eine Diskriminierung nicht explizit in den Programmcode implementiert wird“.

Letztendlich führt ein Softwaresystem, das aus, was im Programmcode steht, wenn der Designprozess der KI-Anwendungen ordnungsgemäß entwickelt wurde. Wenn die Trainingsdaten jedoch fehlerhaft oder unvollständig sind, dann trifft eine antrainierte KI-Anwendung trotz Verwendung der aktuellen Softwarepackages eine nicht gerechtfertigte Entscheidung, weil

> „die abgeleiteten Regeln (…) im Wesentlichen abhängig von der verwendeten Datengrundlage [sind]“ (Zweig 2019a, S. 4).

Anders kann sich eine KI-Anwendung, die auf Künstliche Neuronale Netze basieren (KNN, engl. Artificial Neural Networks, ANN) dann fehlerhaft verhalten, wenn

[3] Auf Open-Source-KI Plattformen sind verschiedene vorgefertigte Algorithmen (bspw. H2O.ai, ClearML, TensorFlow usw.) und gebrauchsfertige Schnittstellen vorhanden, die je nach Lizenzierung kommerziell oder nicht-kommerziell verwendbar.

die Parametermodifizierungen nicht bekannt sind oder der „Implementierungscode nicht einsehbar“ ist (ebd.). Während bei traditionellen Modellen (wie überwachtem, unüberwachtem oder bestärkendem Lernen des ML menschliche Interaktionen durch Programmierende erforderlich sind, um anschließend Algorithmen lernen zu lassen, verfolgt das Verfahren des „Tiefen Lernens (TL)“ (engl. Deep Learning, DL) die Strategie der minimalen manuellen Eingriffe, justiert sich immer wieder neu und entwickelt selbstständig neue Entscheidungsregeln (Alpaydın 2022, S. 329). Vorbild für die Verarbeitungsmethode des TL sind Neuronale Netze des menschlichen Gehirns, die künstlich erzeugt werden (ebd.). Während in den 1970er Jahren KNN aufgrund von geringen Rechenkapazitäten und digitalen Daten weniger Aufmerksamkeit erfuhren, sind sie heute aufgrund von Parallelrechnern und Cloud Servern in der Lage, digital verfügbare große Datensätze, ohne jegliche Vorverarbeitung der Daten zu prozessieren, die als Trainingsdaten dienen (ebd., S. 322 f.). Die Lernalgorithmen des TL benötigen keine aufwendigen Parametermodifizierungen, sondern nur Daten aus denen sie hervorstechende Merkmale durch mehrschichtige neuronale Netze extrahieren können, sodass das System selbstständig aufgrund dieser Extraktion Informationen auf neue Daten anwenden und ableiten kann, um Vorhersagen oder Prognosen zu treffen (ebd., S. 329). Bei solchen KI-Anwendungen handelt es sich um „Black-Boxen“, deren Schlussfolgerungen selbst für die Programmierenden nicht nachvollziehbar sind, weil sie durch hochkomplexe Prozessierungen zu Entscheidungsfindungen kommen. Bemühungen, KI-Anwendungen mit selbstlernenden Algorithmen transparent zu machen stehen vor der großen Herausforderung, komplexe Entscheidungsprozesse sowohl zugänglich als auch verständlich zu machen, weshalb die Interpretierbarkeit und die Vertrauenswürdigkeit in KI-Anwendungen als soziotechnische Systeme von Bedeutung ist (Mittelstadt et al. 2016, S. 6). Die Informationen müssen nicht nur zugänglich sein, sondern auch verständlich, um als transparent zu gelten (ebd.). Durch den prozessual verankerten Einsatz einer KI-Anwendung in Form eines automatisierten Entscheidungssystems, kann die Entscheidung für eine einzelne Person weder rekonstruiert noch bestimmt werden, ob die eigenen personenbezogenen Daten rechtmäßig, vorurteilsfrei sowie sachgerecht verarbeitet wurden (ebd., S. 28). Die Erklärbarkeit umfasst jedoch die technische Funktionsweise einer KI-Anwendung, deren ‚Zweck‘ und ‚Bestimmung‘ (► vgl. Definition Algorithmus von Hill 2016) bleibt jedoch klärungsbedürftig. Insbesondere dann, wenn KI-Anwendungen als Entscheidungssysteme „soziale Konsequenzen für die Teilhabe [von einzelnen Personen] (…)“ (Zweig 2018, S. 33), sowie Auswirkungen auf die Gesellschaft insgesamt, oder Ausschluss- und Einschlussmechanismen der sozialen Ressourcenverteilung wie bspw. Sozialleistungen oder Arbeitsplätze beeinträchtigen können. Daraus können Entscheidungen von KI-Anwendungen resultieren, die nicht

erklärbar, transparent und nachvollziehbar sind. Wenn sich aufgrund von teilhaberelevanten Daten Schlussfolgerungen oder Entscheidungen in KI-Anwendungen ergeben, dann sollte der Zusammenhang zwischen den Daten und der Schlussfolgerung nicht nur zugänglich, sondern auch verständlich sowie offen für eine Überprüfung sein (Mittelstadt et al. 2016, S. 4). Die Offenlegung der KI-Implementierung kann jedoch für gewinnorientierte Unternehmen eine Gefahr sein, wenn bspw. der Implementierungscode einer KI-Anwendung von der Konkurrenz kopiert wird oder sicherheitsrelevante Aspekte einen Angriffspunkt für Institutionen bieten.

Die Tragweite der lernenden Algorithmen in KI-Anwendungen des ML und ihre gesellschaftliche Relevanz für den Einsatz als soziotechnische Systeme erhält eine Schlüsselstellung im Zeitalter der Künstlichen Intelligenz. Wie Tsamados et al. (2021, S. 98) betonen sind Algorithmen ethisch nicht neutral. Welche Kriterien die ethische Neutralität der KI-Anwendungen gewährleisten, ist je nach Fachdisziplin unterschiedlich. Die Informatikerin Zweig (2018) ist der Meinung, dass lernende Algorithmen so einer Konstruktion unterzogen werden können, dass sie weder bei der Urteilsfindung sich von Emotionen leiten lassen, noch „nach Geschlecht oder Herkunft diskriminieren“, und eben aus diesem Grund besäßen KI-Anwendungen ein Potenzial, die die „gesellschaftliche Teilhabe von Menschen“ vergrößern würden, die bisher unter Diskriminierung“ gelitten hätten (Zweig 2018, S. 12). Dass KI-Anwendungen in ihrer bisherigen Karriere als Entscheidungssysteme genau das Gegenteil getan haben, obwohl sie eigentlich zur Verbesserung einer existierenden Problemstellung entwickelt wurden, zeigen die vielfältigen Vorfälle (vgl. Fallbeispiele ► Kapitel 3.3.2). Wie aufgezeigt wurde, bestehen KI-Anwendungen aus mathematischen Konstrukten, die den Gesetzmäßigkeiten von mathematischen Funktionen unterliegen und daher in sich rational arbeiten. Ihre Rationalität hängt von Daten ab, die ihr zur Verfügung gestellt wird. Dazu wird im folgenden Abschnitt der theoretische Bezugsrahmen für Daten in KI-Anwendungen vorgestellt, der den strukturellen Rahmen für die Forschungsfrage und für die systematische Literaturanalyse liefert sowie für die qualitative Inhaltsanalyse.

3. Theoretischer Bezugsrahmen

Im theoretischen Bezugsrahmen werden Daten als zentraler Gegenstand von KI-Anwendungen dargestellt, um im nächsten Schritt einen Bezug zum Feminismus herzustellen. Im letzten Schritt werden der Feminismus und Daten zu ethischen Werten zusammengeführt, um die Kompetenz einer KI-Anwendung herzuleiten.

3.1 Daten als zentraler Gegenstand in KI-Anwendungen

KI-Anwendungen sind datenhungrig und benötigen Daten, um aus diesen im ML zu lernen und Regeln für zukünftige neue Daten zu abstrahieren. Im Folgenden wird anhand der Rolle und Bedeutung von Daten und in KI-Anwendungen ein Bezugsrahmen für die Methodik der qualitativen Inhaltsanalyse skizziert.

3.1.1 Die Rolle der Daten in KI-Anwendungen

KI-Anwendungen benötigen digitale Daten, die durch digitale Artefakte entstehen. Digitale Spuren in Form von Daten werden durch jeden Klick in der digitalen Sphäre hinterlassen, deren Inhalt durch eine KI-Anwendung bspw. von Google, Facebook usw. ausgewertet wird (Otte 2019, S. 444). Diese digitalen Spuren sind sozusagen die Abgase des Informationszeitalters (Schneier 2015, S. 23) und liegen in unverarbeiteter Form als ‚Rohdaten' vor, die von oder über Nutzenden produziert werden (Olteanu et al. 2019, S. 1). Solche Rohdaten sind nicht nur für soziale Medien von wertschöpfendem Interesse (ebd.), sondern auch für Regierungen, die Rohdaten als „die Rohstoffe des 21. Jahrhunderts" bezeichnen (Die Bundesregierung 2016, o.S.).

Während zu Beginn des Internetzeitalters noch rudimentäre Web-Analytics-Systeme im Einsatz waren, die noch zu technischen Analysezwecken eingesetzt wurden, um die Anzahl der Seitenzugriffe und Besuche zu erfassen, speichern und auszuwerten, sind die Analysetechniken (engl. Digital Analytics) heute in der Lage Rohdaten aus der Datensphäre abzuziehen und in interne Datenbanken oder in Drittsysteme von anderen Anbietern zu laden (D'Onofrio et al. 2021, S. 52).[4]

Nach einer Studie (Reinsel et al. 2018, S. 7) werden digitale Daten in einer Datensphäre von drei Kategorien produziert, verarbeitet und gespeichert. Solche Vorgänge in der Datensphäre werden auch Datafizierung genannt. Den Kern (engl. Core) bilden die zentralisierten Rechenzentren in Unternehmen, Organisationen usw. sowie internetbasierte Rechenzentren (engl. Cloud Computing). Im Gegensatz zu Cloud Computing gibt es die dezentralisierten Rechenzentren (engl. Edge Computing) wie bspw. Server vor Ort, Mobilfunktürme und kleinere Rechenzentren, die am Rand des

[4] s. hierfür in ► Kap. 3.3.2 das Beispiel ClearView.

Netzwerks Daten verarbeiten, der sogenannten Edge. Die letzte Kategorie beinhaltet die Endgeräte (engl. Endpoint) wie bspw. Smartphone, PC, Haushaltsgeräte oder industrielle Sensoren in IoT Geräten usw., die eine Verbindung zu einem Computernetzwerk herstellen und dadurch in der Datensphäre digitale Daten erzeugen. Während die globale Datensphäre (engl. Global Datasphere) im Jahr 2018 rund 33 ZB betrug, prognostiziert die Studie bis 2025 einen Anstieg auf 175 ZB.[5] Ein Schlüsselaspekt, der die Datensphäre heute charakterisiert, so die Studie, ist die zunehmend kritische Rolle der Endpoint- und Edge-Kategorien, da hier die sogenannten ‚Rohdaten', entstehen und von besonderem Interesse für Institutionen, Organisationen, Unternehmen und Regierungen sind. Ersteres sammelt digitale Daten, während letzteres diese Daten übernimmt, verarbeitet und an die Core-Kategorie weiterleitet, die wiederrum die Aufgabe hat, die Daten an die Endpoint-Kategorien zu verteilen. Solche Arbeitszyklen ermöglichen latenzabhängige Aktionen wie bspw. bei Suchanfragen. Die Erhebung, Sammlung, Auswertung und Analyse von Daten werden durch KI-Anwendungen verarbeitet, …

> „… die in Bezug auf ein komplexes Ziel auf physischer oder digitaler Ebene handeln, indem sie ihre Umgebung durch Datenerfassung wahrnehmen, die gesammelten strukturierten oder unstrukturierten Daten interpretieren, Schlussfolgerungen daraus ziehen oder die aus diesen Daten abgeleiteten Informationen verarbeiten, und über das bestmögliche Handeln zur Erreichung des vorgegebenen Ziels entscheiden. KI-Systeme (…) sind auch in der Lage, die Auswirkungen ihrer früheren Handlungen auf die Umgebung zu analysieren und ihr Verhalten entsprechend anzupassen" (HEG-KI 2018, S. 6).

Was aus dieser Definition der „Hochrangige[n] Expert*innengruppe für [K]ünstliche Intelligenz" (HEG-KI) implizit umschrieben wird, ist der Begriff *Big Data*. Denn ganz allgemein bezieht sich Big Data auf die Sammlung, Analyse und Nutzung riesiger Mengen digitaler Informationen für die algorithmische Entscheidungsfindung in KI-Anwendungen (McNeely et al. 2022, S. 81). Aus der Definition der hochrangigen Expert*innengruppe der EU lässt sich schlussfolgern, dass Daten eine bedeutende Rolle einnehmen, und in naher Zukunft in Bezug auf ihre Verbreitung, ihren Umfang und ihren Wert noch weiter zunehmen werden (ebd.), weil sie in der heutigen Datensphäre wachsen (Reinsel et al. 2018, S. 7). Infolgedessen werden datengetriebene Big-Data-Tools und -Technologien derzeit entwickelt, um die Echtzeitverarbeitung & -verwaltung großer Mengen verschiedener Daten zu ermöglichen, um Trends und Muster aufzudecken und zu spezifizieren sowie Beziehungen und Verbindungen aufzuzeigen, die für die Entscheidungsfindung, Planung und Forschung von Bedeutung sind (McNeely et al. 2022, S. 81). Verschiedene digitale Daten werden in der Datensphäre als unstrukturierte Daten kontinuierlich durch IoT, soziale Medien und anderen Quellen

[5] One zettabyte is equivalent to a trillion gigabytes (Reinsel et al. 2018, S. 7).

generiert (Hultquist 2022, S. 285). Sie entsprechen keinem vordefinierten Format und sind dadurch vielfältig, wie. bspw. ein Beitrag in den sozialen Medien verschiedenen Formaten entsprechen kann: Logs, html Tags, Videos, Texte, Audio, Bilder usw. (Kadadi et al. 2022, S. 291). Strukturierte Daten hingegen liegen in organisierter Form vor, die wie in einer Excel-Tabelle mit Zeilen und Spalten strukturiert und in Datenbanken abgelegt werden (Alghushairy et al. 2022, S. 341). In der Definition der HEG-KI fehlen noch die semistrukturierten Daten. Diese sind zwar auch unstrukturiert, enthalten jedoch zusätzlich Metadaten, die auf bestimmte Merkmale anderer Daten verweisen (ebd.). Solche Merkmale liefern eine Momentaufnahme mit Kontext über eine Datei, wie z. B. Informationen über die erstellende Person, das Datum, das Thema, den Ort, IP-Adresse, die Zeit und die verwendeten Methoden usw. (Ma 2017, S. 1). Ohne die Metadaten sind Daten (sowie Rohdaten) wertlos, und ihr Potenzial unwiederbringlich verloren (Datenethikkommission 2019, S. 53).

3.1.2 Die Bedeutung der Daten in KI-Anwendungen

Das besondere Interesse liegt an der Erkenntnisgewinnung solcher unstrukturierten, strukturierten oder semi-strukturierten Daten. Hierzu wird der Prozess Data Mining angewandt, der bereits in den 1980er Jahren entstand, als Datenbanken unter menschlicher Kontrolle zur Informationsextraktion untersucht wurden (Prabhu 2022, S. 279). Andere Bezeichnungen für Data Mining sind Wissensextraktion, Informationsentdeckung, Informationssammlung, Datenarchäologie und Datenmusterverarbeitung (ebd.).

Mit Data-Mining-Methoden werden Daten in umfangreichen Datenbeständen entweder zur Prädiktion oder Deskription von menschlichem Verhalten mithilfe von KI-Anwendungen analysiert (D'Onofrio et al. 2021, S. 27). KI-Anwendungen, die zu prädiktiven Zwecken eingesetzt werden, wenden statistische Verfahren (Klassifikation oder Regression) des ML an, um aus historischem Datenbestand zu lernen (ebd.). Dies geschieht im Verfahren des Überwachten Lernens (engl. supervised learning), da Daten, im historischen Datenbestand, mit Informationen zum historischen Verlauf und Resultat einer Situation gekennzeichnet (engl. labeled data sets) werden (ebd.).[6] Mit der Eingabe der gelabelten Daten wird die KI-Anwendung mit ML trainiert. Aus diesen Trainingsdaten lernen die Algorithmen, Zusammenhänge und Abhängigkeiten zu erkennen, und entsprechend einen Regelwerk für den zweiten Algorithmus zu entwickeln (Zweig 2018, S. 13). Die so erlernten Muster, werden durch die antrainierte KI-Anwendung zur

[6] Beispiel: Eine Person möchte bei einem Kreditinstitut ein Darlehen aufnehmen. Das Kreditinstitut kann mithilfe der KI-Anwendung, die mit ML aufgrund der gelabelten Daten gelernt hat, feststellen, wie sich ähnliche Kund*innen verhalten haben. Darauffolgend entscheiden die Algorithmen des ML über die Vergabe oder Kondition des Darlehens aufgrund von historischem Datenbestand.

Erstellung für korrekte Prognosen, auf neue und unbekannte Daten eingesetzt (Alpaydın 2016, S. 39)

Die deskriptive Verfahrensweise der Data-Mining-Methode wird angewandt, um aus riesigen Datenmengen, die unstrukturiert in der Datensphäre vorliegen, neue Informationen bzw. Erkenntnisse zu extrahieren. Aufgrund der Menge sind die Daten ungelabelt, also ohne irgendwelche Informationen gekennzeichnet und unstrukturiert vorhanden (D'Onofrio et al. 2021, S. 27). Das Ziel des unüberwachten ML (engl. unsupervised maschine learning) ist, Muster in den unstrukturierten Daten zu erkennen. Die Wissensextraktion kann zur Analyse von Anomalie Erkennung verwendet werden (bspw. bei Betrugsversuchen oder terroristischen Aktivitäten, die vom Normalzustand abweichen), oder zur Dichteschätzung bzw. Häufungen von Datenaufkommen (Clusteranalyse wird bspw. in Marktforschungen eingesetzt, die eine heterogene Gruppe von Objekten in homogene Untergruppen aufteilen), oder als Assoziationsanalysen (bspw. zur Warenkorbanalyse bei Onlineversandhändlern) (Alpaydın 2022, 5 ff.). Das Risiko hierbei ist, dass neue und unbekannte Muster in den Daten mit personenbezogenen Informationen, zu ungeahnten Ungleichbehandlungen führen können (► vgl. Fallbeispiele im Kap. 3.2.3).

Solche Data-Mining-Methoden, die mit personenbezogenen Daten Wissen aus großen Datenbeständen generieren, werden Profiling genannt. Gem. der Datenschutz Grundverordnung (DSGVO) im Artikel 4, Nr. 4 ist …

> „… Profiling jede Art der automatisierten Verarbeitung personenbezogener Daten, die darin besteht, dass diese personenbezogenen Daten verwendet werden, um bestimmte persönliche Aspekte, die sich auf eine natürliche Person beziehen, zu bewerten, insbesondere um Aspekte bezüglich Arbeitsleistung, wirtschaftliche Lage, Gesundheit, persönliche Vorlieben, Interessen, Zuverlässigkeit, Verhalten, Aufenthaltsort oder Ortswechsel dieser natürlichen Person zu analysieren oder vorherzusagen“ (EU 2016, S. 33).

KI-Anwendungen mit ML, die durch Daten lernen, spiegeln die historischen und aktuellen Werte einer Gesellschaft wider. Daten sind keine Rohstoffe, die in unverarbeiteter Form, wie in der Natur vorkommen, sondern sind etwas Technisches, weil sie Messwerte sind und technisch erhoben werden. Sie werden digital in Anerkennungsprozessen gestaltet, die auf persönlichen Informationen eines Subjekts mit Rechtsansprüchen beruhen. Durch die Verarbeitung der Daten von Individuen mit den dazugehörigen Entscheidungsalgorithmen in KI-Anwendungen des ML können u.a. kulturelle Unterschiede oder andere Differenzkategorien entstehen, die sich diskriminierend auswirken und in gesellschaftlichen Strukturen tief verankern und bedingt disponibel sind. Somit können KI-Anwendungen Einflussfaktoren auf Teilhabechancen haben, wenn es bspw. um Ressourcenverteilung handelt. Laut Zweig (2018, S. 33) können im Algorithmendesign Fehler vorkommen, jedoch sind diese

schnell lösbar, wenn zum einen der Quellcode zugänglich ist, und zum anderen der Einsatzzweck von Algorithmen zur Lösung eines Problems bekannt ist. Die Eingabe- bzw. Trainingsdaten werden jedoch durch unterschiedliche Akteure wie staatliche, wirtschaftliche und wissenschaftliche Institutionen oder NGOs[7] gesammelt und von Programmierenden ausgewählt und in KI-Anwendung mit lernenden Komponenten eingesetzt (ebd., S.17 f.). Die kognitive Leistung des ML liegt dann in der Reflexion der gesellschaftlichen Machtstrukturen, die sich in den Beispiel- oder Trainingsdaten durch Mustererkennung entdecken lässt. Insofern beeinflusst die Datenqualität maßgeblich die korrekte Arbeitsweise der KI-Anwendungen, wenn diese als soziotechnische Systeme agieren. Die Datenqualität ist der Grad, in dem die Daten die Anforderungen des beabsichtigten Konstruktionszwecks und -bestimmung einer KI-Anwendung erfüllen. Für das BSI (2019a) ist eine hohe Datenqualität gegeben, wenn Daten aus vertrauenswürdigen Quellen stammen und korrekt kommentiert und angemessen geschützt sind (BSI 2021a, S. 9). Dies würde jedoch bedeuten, dass KI-Anwendungen sicher sind, wenn die Trainingsdaten in strukturierter Form im ML eingesetzt werden, was jedoch dem Big Data Verfahren widerspricht, da das besondere Interesse gerade an der Wissensextraktionen aus Rohdaten besteht. Genauso wie Datenqualität ist jedoch auch das Datenmanagement wichtig, welches die Schritte von der Erhebung und Sammlung sowie Speicherung und Verarbeitung bis hin zur Archivierung oder Löschung der Daten umfasst (ebd.). So werden in verschiedenen Fachdisziplinen über den Einsatz von personenbezogenen Daten diskutiert, positionsbezogen ausgehandelt, Mechanismen für gesellschaftliche Dynamiken entwickelt, prosperierende Wirtschafts- und Erwerbsbereiche unterliegen dem technischen Wandel und erfahren Disruptionen, strukturelle und institutionelle Ausschlussmechanismen, die in den erzeugten Daten wiederzufinden sind, werden von der KI-Anwendung übernommen und wirken wie ein Verstärker auf einzelne Individuen oder soziale Gruppen. So könnte die Liste noch weitergeführt werden. Die zivilgesellschaftlichen Diskurse, die ökonomiegetrieben sind, beinhalten die sozialen und technischen Verschränkungen rund um KI-Anwendung mit allerhand Maßnahmen, Entwicklungen, Lösungs- und Handlungsansätze. Dabei erlebt der ethischer Anspruch an KI-Anwendungen ein Nischendasein. Hochwertige Daten, die unvoreingenommen und diskriminierungsfrei sind, werden als Mangelware angesehen. Versuche, Daten so aufzubereiten, dass sie in KI-Anwendungen mit geringer Fehlerquote eingesetzt werden können, verfolgen die Strategie der Ausbeutung. So sitzen Menschen an Computern, um Bildklassifizierungen für ein paar Cent zu erledigen oder auf Crowdsourcing-Plattformen sind Anbietende, die ‚Crowdworker' für die sogenannten ‚Human Intelligence Tasks' einsetzen, um Dateien zu kennzeichnen

[7] NGO ist eine Abk. für Non-Governmental Organisation - Nichtregierungsorganisation

(labeln). Solche Crowdworker verrichten ihre Arbeit jedoch unter prekären Arbeitsverhältnissen (Dierks 2017, o.S.). ImageNet, eine Open-Source Datenbank, die knapp 14 Millionen Bilder aus der Datensphäre zu Forschungszwecken für ML bereitstellt, nutzt solche Crowdworker, um Bilder, die im Internet vorhanden sind, zu labeln. Ziel der Plattform ist, das Wissen in der realen Welt in digitalen Daten abzubilden. Einem online Artikel zufolge haben ML, die die Datensätze von ImageNet nutzen, eine Fehlerquote von zwei Prozent, wenn sie Bilder klassifizieren bzw. identifizieren – das kontextabhängige Verständnis fehlt jedoch gänzlich (Gershgorn 2017, o.S.).

KI-Anwendungen mit ML, die personenbezogene Daten nutzen, zeigen ungerechtes Verhalten auf, wenn die Qualität der Daten schlecht, fehlerhaft oder nicht vollständig ist. Im Folgenden wird der theoretische Rahmen für den Feminismus gelegt, der den Gerechtigkeitsaspekt thematisiert.

3.2 Der Feminismus

Um die feministische Perspektive zu erhalten bzw. zu definieren, ist zunächst die Auseinandersetzung mit dem Terminus Feminismus und feministischen Theorien unabdingbar. Infolgedessen wird zunächst eine Definition vorgenommen. Entlang der Entstehung und Entwicklung des Feminismus werden über die feministischen Bewegungen ein allgemeiner Überblick gegeben sowie Konzepte und Theorien dargestellt, die der wissenschaftlichen Analyse der vorliegenden Arbeit konkret dienen.

3.2.1 Die Definition von Feminismus

Laut Duden ist die semantische Bedeutung des Begriffs Feminismus ein „Oberbegriff für verschiedene Strömungen, die sich für die Gleichberechtigung, Selbstbestimmung und Freiheit aller Geschlechter, v.a. von Frauen, und gegen Sexismus einsetzen, z.B. durch das Anstreben einer grundlegenden Veränderung gesellschaftlicher Normen (z. B. der traditionellen Rollenverteilung) und der patriarchalischen Kultur“ (DUDEN 2022, o.S.). In der Encyclopaedia Britannica hingegen wird der Feminismus folgendermaßen definiert:

> „Feminism, [is] the belief in social, economic, and political equality of the sexes” (Burkett et al. 2021, o.S.).

Lenz (2018, o.S.) wiederrum konstatiert, dass das „Grundanliegen aller feministischen Strömungen[8] die Selbstbestimmung, Freiheit und Gleichheit für alle Menschen sind, die im öffentlichen wie auch im persönlichen Leben verwirklicht werden

[8] Liberaler Feminismus, Black Feminism (Intersectional Feminism), Womanism, islamischer Feminismus, afrikanischer Feminismus, Queer Feminismus, postkolonialer Feminismus (Third World Feminism), sozialistischer Feminismus, Differenzfeminismus (auch in radikaler Form), Öko Feminismus, diskurstheoretischer Feminismus, konservativer Feminismus, transformativer Feminismus usw.) und ist dadurch sehr vielfältig aufgestellt.

soll“. Unterschiedliche feministische Strömungen definieren Feminismus aus der eigenen Perspektive. Lenz (2019, S. 2) fordert, dass durch die Vielfalt „eher von Feminismen“ gesprochen werden sollte. Wohingegen Hark (2007, S. 13) für die Terminusverwendung im Singular plädiert, damit die feministische Theorie sich etablieren kann. Dabei geht es nicht um die Unkenntlichmachung von „komplexer Vielstimmigkeit feministischer Perspektiven und Positionen“, sondern um „Feminismus insgesamt komplexer zu reformulieren“, damit „eigene Konzepte von Denken, Erkennen und Wissen“ in der männlich (*weiß*)[9] geprägten Wissensproduktion, etabliert werden (ebd. S. 9-11).

Wenn der Begriff Feminismus etymologisch zerlegt wird, so ist die Wortherkunft aus dem Französischen und steht für *féminisme*, wobei der Begriff vom Lateinischen *femina* abgeleitet wird, dessen Bedeutung in der deutschen Sprache ‚Frau' ist. Der Suffix -ismus ist ein Wortanhängsel und kommt aus dem Lateinischen, der vielfach eine negative Konnotation in gesellschaftlichen Diskursen hervorruft. In diesem Zusammenhang kämpft der Feminismus gegen den Exklusionsmechanismus in gesellschaftlichen Strukturen. Oftmals wird mit dem Begriff die Entmachtung von Männern verstanden. Frauen, die sich aktiv für ihre Rechte einsetzen, wird „übertriebene Männerfeindlichkeit und Männerhass unterstellt und abschätzig als weibliche Form des Sexismus interpretiert“ (Strauß 1989, S. 107). Dabei ist in Anbetracht der Historie die konsequente Entmachtung der Frauen festzustellen. Verglichen mit Männern sind Frauen demokratische Neulinge, da sie demokratische Teilhabe seit durchschnittlich etwa einem Jahrhundert praktizieren (Holland-Cunz 2019, S. 1). Feminismus ist also der Hinweis auf „Demokratisierungsdefizite“ (ebd., S. 2).

Auch Männer hatten nicht von Anfang an alle Rechte, weil sie männlich waren. Abhängig von ökonomischem Status wurde ihnen die gesellschaftliche und politische Teilhabe gestattet. Während der Französischen Revolution (1789 bis 1799) hatte das allgemeine Männerwahlrecht eine symbolische Bedeutung. Knapp 50 Jahre später galt das Zensuswahlrecht in ganz Europa immer noch, das je nach Einkommen und Steuerabgaben die Wahlbeteiligung der Männer vorsah. Das allgemeine Wahlrecht für Männer galt als politische Innovation, das mit Hürden und Konflikten, von der Ersteinführung (Ende des 17. Jahrhunderts) bis zur Verwirklichung (Mitte des 18. Jahrhunderts), zuerst in den USA, dann in Frankreich und in der Schweiz eingeführt wurde, die jedoch bei der Einführung des Frauenwahlrechts das Schlusslicht bildeten. In den Vereinigten Staaten zeigt sich jedoch, dass sich die „Dominanz *weißer*,

[9] Mit der Bezeichnung *weiß* wird der „Status eines *unmarkierten Markierers* und seine *unsichtbar herrschende Normalität*“ verdeutlicht (Arndt 2020, S. 21).

protestantischer Männer“ nicht nur auf die Ungleichberechtigung der *weißen* Frauen auswirkte, sondern auch auf viele US-Bürger mit afroamerikanische, indigene und lateinamerikanische Ethnizität, die mit Sonderbedingungen vom Wahlrecht bis zum Jahr 1965 ausgeschlossen wurden (Raschke 2020, 395 ff.). Viel subtiler treten solche Diskriminierungsformen zu Tage, wenn strukturelle Ausschlussmechanismen durch KI-Anwendungen aufgedeckt werden. Bei den Wahlen 2020 wurde die indigene Bevölkerung in den USA als „Something Else“ bezeichnet, weil Nachrichtensender, unreflektiert die Hochrechnung der KI-Anwendungen übernahmen und ausstrahlten (Forester 2020, o.S.).

Laut Gerhard (2020) wird oftmals der Begriff Feminismus mit Frauenbewegungen synonym verwendet. Das gemeinsame Ziel von Frauenbewegungen und Feminismus ist: „Frauen in allen Lebensbereichen, in Staat, Gesellschaft und Kultur und vor allem auch in der Privatsphäre, gleiche Rechte und Freiheiten sowie gleiche Teilhabe an politischer Macht und gesellschaftlichen Ressourcen zu verschaffen“. Mit Frauenbewegungen ist das gemeinsame soziale Handeln von Frauen gemeint, die einen sozialen Wandel im Kontext der Geschlechterverhältnisse anstreben, wobei der Feminismus noch eine weitergehende Bedeutung hat (Gerhard 2020, S. 7).

Frauenbewegungen setzten sich für Gleichberechtigung ein und sind maßgeblich an der Umsetzung des Wahlrechts beteiligt (ebd., S. 9). Die „großen westlichen Demokratien führten erst nach dem Ersten, oder – in den katholischen Ländern – nach dem Zweiten Weltkrieg“ das Frauenwahlrecht ein (Raschke 2020, S. 74). Frauenwahlrecht ist insofern wichtig, da mit dem Wahlrecht das Mitbestimmen und Mitgestalten der eigenen Rechte sowie die der gesellschaftlichen Strukturen garantiert. Umso interessanter ist es, wenn vom biologisch geschlechtlichen Standpunkt aus betrachtet wird, dass die Hälfte einer bestehenden Gesellschaft vom aktiven und passiven Wahlrecht systematisch ausgeschlossen wurde, was eigentlich ein demokratisches Grundprinzip ist. Dabei ist zu beobachten, dass die Länder, die das Frauenwahlrecht nur dann einführten, wenn es um „übergreifende Mobilisierung und Legitimierung“ (Raschke 2020, S. 488) einer politischen Agenda ging. Folglich werden Machtstrukturen nur dann verändert, wenn es um Vorteile der eigenen (männlichen) Machterhaltung geht oder eben diese gefährdet sind. Daher waren laut Hooks (2000, S. 4) *weiße* Männer bereit, die Rechte der Frauen zu berücksichtigen, wenn die Gewährung dieser Rechte der Aufrechterhaltung der *weißen* Vorherrschaft dienen konnte.

Frauenbewegungen kritisieren die Unverhältnismäßigkeiten der Machterhaltung, die in einem Gesellschaftskonzept enthalten sind, wohingegen der Feminismus laut Gerhard (2020, 7 f.) „nicht nur einzelne Anliegen verfolgt, sondern die Gesamtheit gesellschaftlicher Verhältnisse im Blick hat“. Dementsprechend verhandeln

Frauenbewegungen die Emanzipation von Frauen im sozialen Gefüge, wohingegen der Feminismus „einen grundlegenden Wandel der sozialen und symbolischen Ordnung […] anstrebt und gleichzeitig Deutungen und Argumente zu ihrer Kritik anbietet“ (ebd., S. 8). Insofern werden im Feminismus Aushandlungsprozesse von „unterschiedlichen Geschlechterkonzepten und Gesellschaftstheorien sowie auf gesellschaftliche Grundfragen wie die Selbstbestimmung über Körper, Sexualitäten und Gebären, die Gleichheit in der Arbeit und der Politik oder den Kampf gegen Gewalt und Krieg“ (Lenz 2019, S. 2) verhandelt.

3.2.2 Feminismus mit intersektionaler Perspektive

Feminismus ist wie Hooks (2000, S. x) sagt, für alle. Eine feministische KI-Anwendung ist demnach auch für alle und berücksichtigt die unterschiedlichen Lebensrealitäten der Menschen in ihrem Entscheidungsprozess. Um die Lebensrealitäten von Menschen zu verstehen, bietet die Perspektive der Intersektionalität einen Analyserahmen für KI-Anwendungen. In Worten von Dzodan (2011, o.S.) ist Feminismus entweder intersektional „… or it will be bullshit!“. Der Schwarze[10] Feminismus mit intersektionaler Perspektive bietet das Verständnis über Lebensrealitäten von Menschen. Die Lebenserfahrungen von Schwarzen Frauen, die aufgrund von multidimensionalen Diskriminierungsmechanismen als marginalisierte Gruppen benachteiligt werden, haben Analyseinstrumente entwickelt, die die verkürzte Sichtweise vom *weißen* Feminismus erweitern. Während der *weiße* Feminismus aufgrund von Genderungerechtigkeit Theorien gegenüber der „nicht-rassifizierten, nicht-vergeschlechtlichten Objektivität maskierte, *weiße*, männliche Subjektivität“ entwickelt, um vermeintlich für die Belange aller Frauen zu sprechen, werden „Frauen of Color[11] in der Tat [nicht nur] übersehen, sondern ihr Ausschluss wird verstärkt“, weil „die Rolle von *Race*[12] oft übersehen“ wird (Crenshaw 2019, 167 f.). Crenshaw (2019) hat durch das Zusammenwirken von *Race* und Gender[13] aufgezeigt, dass der „Diskriminierungsdiskurs um die Intersektion“ erweitert werden muss, um die Diskriminierungskonzepte zu verstehen (ebd., S. 186). Erst wenn die Perspektiven von marginalisierten Gruppen miteinbezogen werden, so Crenshaw (2019), kann der Demokratisierungsdefizit aufgehoben werden.

Mit der Frage „Ain’t I a Woman?“ dekonstruierte die versklavte Freiheitskämpferin Sojourner Truth (1797-1883) (Truth 2019, S. 18) nicht nur die Lesart der damaligen

[10] Die Großschreibung des Adjektivs ist eine Widerstandsform gegenüber rassistischen Konstruktionen (Arndt 2020, S. 21).
[11] Frauen of Color ist eine Selbstbezeichnung von Frauen, die Rassismus erfahren.
[12] Mit dem Begriff *Race* wird verdeutlicht, dass es keine Rassen gibt und eine Erfindung des Rassismus ist.
[13] Während mit Gender (engl.) die soziale Geschlechterrollen bezeichnet werden, ist mit sex (engl.) das biologische Geschlecht gemeint.

politischen Diskurse über Frauen, die als homogene Gruppe galten, sondern machte zum ersten Mal durch ihre eigene Lebenserfahrung darauf aufmerksam, dass sie als Schwarze Frau mehrfache Ungerechtigkeiten durch Unterdrückung erlebt hat (Kelly 2019, S. 10). Dies ist laut Kelly (2019, ebd.) der erste Hinweis auf Intersektion, denn Truth kritisiert im Jahr 1851 nicht nur die Präsenz von Rassismus und Klassenunterdrückung in den Frauenbewegungen, die um das Wahlrecht kämpften, sondern auch die Genderhierarchien beim Stimmrecht sowie die sexistische Diskriminierung innerhalb ihrer eigenen Community, womit sie auf die spezifischen Diskriminierungserfahrungen von Schwarzen Frauen hinweist. Denn diese gleichen weder den Diskriminierungserfahrungen der Schwarzen Männer noch denen der *weißen* Frauen und sind auch nicht in der Summe dieser Erfahrungen zu verhandeln (ebd.).

Die juristische Dimension der Mehrfachdiskriminierungserfahrungen von Schwarzen Frauen wurden erst durch Kimberlé Crenshaw (2019) im Jahr 1989 aufgezeigt, als sie mithilfe der Metapher einer Straßenkreuzung (engl. intersection) das Zusammenwirken von *Race* und Gender verdeutlichte, womit der Begriff Intersektionalität entstand. Der eindimensionale Fokus auf eine Diskriminierungsform führte bei Gerichtsentscheidungen dazu, dass Schwarze Frauen ihre Rechte vor dem Gericht nicht geltend machen konnten, weil sie entweder in die Analysekategorie von Schwarzen Männern aufgenommen wurden oder in die der *weißen* Frauen. Schwarze Frauen wurden nicht nur aufgrund von *Race*, sondern auch wegen Gender und Class[14] diskriminiert, die sozusagen genau in der Mitte einer Straßenkreuzung verortet sind und zu verzerrten Urteilsfindung der Gerichte führten. Mit Intersektionalität beschreibt Crenshaw das komplexe Phänomen von mehrdimensionalen Diskriminierungserfahrungen, die sich gegenseitig beeinflussen und in ihrer Verschränkung insbesondere die Benachteiligung von Schwarzen Frauen ausblenden sowie zu problematischen Effekten führen (Crenshaw 2019, 145 ff.).

3.2.3 KI-Anwendungen und Feminismus

Laut der Expertise von KITE (KI Thinktank Female Entrepreneurship) haben „feministische Strömungen (…) langjährige Erfahrung mit der Benennung und Bekämpfung diskriminierender Strukturen“, die aufgezeigt haben, dass KI-Anwendungen nicht neutral sind, weil sie „ein Resultat aus herrschenden Machtstrukturen“ sind (Fritsch et al. 2021, S. 2). Diese reproduzierten Machtstrukturen werden derzeit lediglich von einem kleinen Teil der Menschheit entwickelt, die bereits den Alltag von fast allen anderen Menschen verändern (Floridi et al. 2021, S. 12).

[14] Class oder „Klassismus ist Ausbeutung, Marginalisierung, Gewalt, Macht und Kulturimperialismus aufgrund der sozialen Herkunft oder Position.“ (Kemper 2016, S. 6)

Intersektionale feministische Positionen fordern Inklusion und Diversität in den Entwicklungs-, Implementierungs- und Anwendungsphase, um Diskriminierung durch KI-Anwendungen entgegenzuwirken (Fritsch et al. 2021, S. 2). So ist Inklusion zu einem Markenzeichen der Bemühungen von akademischen Konferenzen, Forschungszentren und großen Technologieunternehmen geworden, die zum einen dadurch die Datenwissenschaft[15] (engl. Data Science) und die KI-Technologien verbessern und zum anderen ihre Schäden minimieren wollen (Hoffmann 2021, 3545 f.). Denn Inklusion verspricht die Voreingenommenheit und Diskriminierung in der Datenwissenschaft abzumildern, wenn bspw. die Trainingsdaten nicht ausreichend repräsentativ sind, kann Inklusion von verschiedenen Menschen mit unterschiedlichem Background die Verzerrung in den Datensätzen erkennen und verhindern (ebd., S. 3548). Die Realität jedoch spiegelt eine andere Tatsache, denn „die KI-Forschung und ihre Entwicklungsbereiche sind noch nicht divers genug“ (Fritsch et al. 2021, S. 2). Mit Diversität wird aber in den Diskursen oft der Frauenanteil in der IT-Branche thematisiert, der nach wie vor gering ist und rund 20 Prozent der Beschäftigten in den Big-Tech Konzernen (wie Apple, Google, Facebook und Amazon) ausmacht (Nier 2018, o.S.). Das Diversitätskonzept (engl. diversity concept) umfasst jedoch nicht nur die Genderproblematik, wenn der Hintergrund seiner Entstehungsgeschichte betrachtet wird. Der Ursprung des Diversitätskonzepts lässt sich in der Schwarzen US-amerikanischen Bürger*innenrechtsbewegung finden, welches zur Förderung benachteiligter Gruppen eingesetzt wurde (Kelly 2021, S. 59). Im deutschen Kontext wird das Diversitätskonzept sogar gesetzlich geregelt, jedoch beschränkt sich die Diversität auf „die Aspekte wie bspw. Alter, Geschlecht, Bildungs- oder Berufshintergrund“ (vgl. §289f Abs. 2 Nr. 6 HGB). Die Multiperspektivität (Kelly 2021, S. 58), die in den Datensätzen fehlen, werden durch Gesetze strukturell verankert, die sich in den Trainingsdaten wiederfinden. Der ethische Umgang mit Daten ist eine Herausforderung, die die EU als ‚Marktlücke’ kommuniziert, und deshalb auf der Suche nach Lösungen ist, um „vertrauenswürdige“ KI-Anwendungen zu entwickeln (Delcker 2019, o.S.).

Floridi et al. (2021) konstatieren, dass Ethik nicht die Domäne eines einzelnen Kontinents oder einer einzelnen Kultur ist. Den Publizierenden nach, habe jedes Unternehmen, jede Regierungsbehörde und jede akademische Einrichtung, die KI konzipiert, entwickelt oder einsetzt, die Pflicht, dies im Einklang mit einem ethischen Rahmen zu tun, um ein geografisch, kulturell und sozial vielfältigeres Spektrum von Perspektiven einzubeziehen (Floridi et al. 2021, S. 14). Ohne die Einbeziehung der Perspektiven wird Kulturimperialismus ausgeübt (s.o. für Definition Klassismus).

[15] Datenwissenschaft (engl. Data Science) bezeichnet die Wissensextraktion aus Daten.

In der Auseinandersetzung des ethischen Rahmens für KI-Anwendungen als soziotechnisches System muss dies „in Kontext mit größeren Zusammenhänge[n] wie Patriarchat, Kapitalismus oder Kolonialismus gebracht werden“ (Fritsch et al. 2021, S. 2). Hier haben die EU und auch Deutschland noch Nachholbedarf, wenn bspw. keine klare Definition vom strukturellem Rassismus für die Rechtsprechung verbindlich vorhanden ist (Kelly 2021, S. 42).

3.2.4 Feministisch intersektionale Perspektiven für eine ‚Erklärbare KI-Anwendung‘

Der Feminismus übt Kritik auf die Exklusionsmechanismen von gesellschaftlichen und sozialen Strukturen aus und zeigt die Demokratisierungsdefizite auf, die sich auf die androzentrisch geprägte Hegemonie beziehen. Dies geschieht nicht nur aus der Geschlechterperspektive, sondern auch auf der Ebene der politischen Gesellschaftskonzeption. Solche Gesellschaftskonzeptionen bestehen aus Exklusions- und Inklusionsmechanismen, die Individuen aufgrund ihrer Zugehörigkeit zu einer sozialen Gruppe entweder bevorzugen oder benachteiligen. KI-Anwendungen mit ML, die mit Daten trainiert werden, die aus solchen Gesellschaftskonzeptionen entstanden sind, haben in der Vergangenheit aufgezeigt (► vgl. Kap. 3.3.3), dass feministisch intersektionale Perspektiven für eine ‚Erklärbare KI-Anwendung‘ von evidenter Tatsache sind. KI-Anwendungen, deren Vorhersagen oder Prognosen nicht nachvollziehbar sind, werden daher Black-Box genannt (► vgl. Kap. 2.3). Mit Explainable AI (XAI) hingegen sollen die Entscheidungsvorgänge der KI-Anwendungen mit ML transparent sein. Um KI-Anwendungen ethisch und rechtlich bewerten zu können, ob diese transparent sind, „ist es essenziell, dass ausreichend Informationen über [deren] Reichweite, Funktionsweise, Datengrundlage und Datenauswertung zur Verfügung stehen“ (Datenethikkommission 2019, S. 169). Insbesondere dann, wenn Menschen durch KI-Anwendungen bewertet werden oder über sie entschieden wird. XAI-Anwendungen, die auf Fairness achten, versuchen Methoden bereitzustellen, die bei der Verarbeitung von personenbezogenen Daten fair sind und nicht diskriminieren (Friedler et al. 2019, S. 3), also gerecht Entscheidungen treffen. Um Bias entgegenzuwirken, gibt es bei technisch orientierten Lösungsansätzen den Versuch, Daten vor Eingabe in die KI-Anwendung mit ML so aufzubereiten, dass das Ergebnis eines auf diese Daten angewendeten selbstlernenden Algorithmus fair ist (engl. Preprocessing algorithms). Der Vorschlag hierbei ist alle personenbezogenen sensiblen Informationen, die zu einer unfairen Entscheidung einer KI-Anwendung führen, aus dem Datenbestand zu entfernen oder hinzuzufügen. Ersteres soll gemacht werden, wenn die Trainingsdaten selbst eine Diskriminierung beinhalten, weil sie die strukturellen Diskriminierungen der realen Welt abbilden. Dieses Prinzip „fairness through unawareness“ führt jedoch dazu, dass KI-

Anwendungen trotz Fehlinformationen bzw. Wissen in der Lage sind, durch Korrelationen von Proxies (stellvertretende Merkmale) eine ungerechtfertigte Lösung zu finden, die zu ungerechten Entscheidungen oder Bewertungen führen können (Ammon et al. 2021, S. 38). Letzteres soll gemacht werden, wenn eine Gruppe von Menschen unterrepräsentiert und aufgrund dessen diskriminiert werden (ebd.). Weitere technische Lösungsansätze sind bereits in ► Kap. 2.3 dargestellt. Hinzu kommen noch über 70 technische Fairnessmaße vom Unternehmen IBM, welches mit seinem AI Fairness 360 - Open Source Toolkit, die Diskriminierungen und Bias in KI-Anwendungen eindämmen will (IBM Research Trusted AI 2018, o.S.). Was aber genau die ethischen Werte einer KI-Anwendung sind und welche Herausforderungen sich ergeben, wird im Folgenden aufgezeigt.

3.3 Ethische Werte einer KI-Anwendung

Die Datenethikkommission (2019) ist der Meinung, dass ethische Grundsatzfragen wie Teilhabe, Fairness, Gleichbehandlung, Selbstbestimmung und Transparenz bzgl. der Datenverwendung nicht an die Technik delegiert, oder gar den Entwickelnden von KI-Anwendungen überlassen werden sollten, sondern vielmehr dem Konzept des „Ethics by Design“ folgen sollten. Dieses Konzept, so die Datenethikkommission, könnte bereits bei der Entwicklungs- und Entstehungsphase von einer KI-Anwendung angewandt werden, um kontextspezifisch ethische Prinzipien mitzudenken sowie ggf. Betroffene einzubeziehen (Datenethikkommission 2019, S. 74).

So hebt sich hervor, dass allgemeingültige Handlungsanweisungen für Programmierende *nur* technischer Natur sein können. Sie können die Fairness einer KI-Anwendung insoweit umsetzen, wie eine mathematische Funktion dies zulässt. Mathematische Funktionen benötigen Messwerte, die in Zahlen als Fairnessmaße ausgedrückt werden können. Fairnessmaße gegen Diskriminierungen sind, wie später in den Fallbeispielen aufgezeigt wird, jedoch noch in der Entwicklungs- und Aushandlungsphase. Dies geht auch aus der Datenstrategie (Die Bundesregierung 2021, S. 7) hervor:

> „Bei der Nutzung von Daten ist nicht alles, was technisch möglich ist, auch ethisch vertretbar und politisch wünschenswert. Die mit der Verarbeitung von großen Datenmengen verbundenen Möglichkeiten, über → Profiling und Scoring Verhalten zu prognostizieren und zu steuern sowie Präferenzen zu beeinflussen, müssen kritisch hinterfragt und gegebenenfalls begrenzt werden. Datennutzung darf nicht zu einer sozialen oder politischen Polarisierung führen. Datenrecht und ethische Grundsätze sind keine Bremse, sondern wichtig für den Schutz der Grundrechte und eine verantwortungsvolle Datennutzung“.

Für die verantwortungsvolle Datennutzung in KI-Anwendungen mit ML würde dies bedeuten, dass soziale oder politische Ungerechtigkeiten, Einschluss- und Ausschlussmechanismen sowie die Unterdrückungs- und Manipulationsmechanismen

verhindert werden, wenn Methoden zur Eindämmung von Diskriminierung und Bias vorhanden wären.

Scherr (2010) charakterisiert Diskriminierung als eine Unterscheidungspraxis, „mit der ‚die Normalen' von denjenigen unterschieden werden, ‚die in unerwünschter Weise anders' sind" (Scherr 2010, S. 43). Nach Kelly (2021, S. 39) werden solche Unterscheidungspraxen im Grundgesetz durch Artikel 3 Absatz 3 mitgedacht, da das fundamentale Verbot der ‚*weißen* Vorherrschaft' sich damit ableiten lässt, die als die ‚Normalen' von Scherr (2010) bezeichnet wird. Die ‚in unerwünschter Weise anders' sind, sollen durch Artikel 3 Absatz 3 im Grundgesetz geschützt werden:

> *„Niemand darf wegen seines Geschlechtes, seiner Abstammung, seiner Rasse, seiner Sprache, seiner Heimat und Herkunft, seines Glaubens, seiner religiösen oder politischen Anschauungen benachteiligt oder bevorzugt werden. Niemand darf wegen seiner Behinderung benachteiligt werden."*

Liegt eine ungerechtfertigte Gleich- oder Ungleichbehandlung (Beck et al. 2019, S. 7) aufgrund der o.g. Diskriminierungsmerkmale vor, dann hat Diskriminierung weitreichende Auswirkungen auf die Lebensbedingungen und Lebenschancen (Scherr 2010, S. 35). Die Auswirkungen betreffen Ausschluss- und Einschlussmechanismen zu Ressourcen der Gesellschaft wie bspw. Arbeit, Aufenthaltsstatus, Bildung, Eigentum, soziale Leistungen usw. von sozialen Gruppen oder deren Angehörige, wenn der Zugang durch Diskriminierung geregelt wird (ebd.). Solche Ausschluss- und Einschlussmechanismen entstehen aufgrund von Vorurteilen und Stereotypisierung (Mehrabi et al. 2021, S. 10) gegenüber sozialen Gruppen oder deren Angehörige.

3.3.1 Bias-Effekte in KI-Anwendungen

Ähnlich wie die Diskriminierung ist auch Bias eine Form der Ungerechtigkeit (Mehrabi et al. 2021, S. 10). Das Wort Bias stammt aus dem engl. ab und bedeutet Verzerrung, Voreingenommenheit, Vorurteile, Einseitigkeit, Abweichung, Bevorzugung (s. deepl Übersetzungstool). Während menschliche Vorurteile und Stereotypisierungen zu diskriminierender Ungerechtigkeit führen, aufgrund von nicht veränderbaren und sensiblen Merkmalen einer Person oder sozialen Gruppen, beruhen Bias-Effekte bei KI-Anwendungen auf Daten (Sammlung bzw. Erhebung, Speicherung und Verarbeitung) und Stichprobenbildung in statistischen Modellierungen der Algorithmen sowie Messwerten von mathematischen Funktionen (ebd.). Somit liegt es in der Natur des ML, beim Verarbeiten der Daten, diese gruppenbezogen statistisch zu sortieren bzw. zu strukturieren oder zu klassifizieren oder einen Scoring-Wert zuzuweisen, um aus den hervorstechenden Merkmalen des Datenbestands Wissen zu extrahieren. Bias-Effekte können nicht nur technischer Natur sein, sondern auch durch Menschen entstehen, deren Vorurteile in die Entwicklungs-, Implementierungs-, oder Anwendungsphase einer

KI-Anwendung hineinfließen (Zweig 2018, S. 33). Wenn KI-Anwendungen diskriminierend handeln, weil sie auf Daten trainiert werden, die Vorurteile beinhalten, dann können die zugrundeliegenden Trainingsdaten ihren Ursprung in der Gesellschaft insgesamt, in Subkulturen und in formellen oder informellen, privaten oder öffentlichen Organisationen und Institutionen haben (engl. pre-existing Bias) (Friedman et al. 1996, S. 333). Lösungen, die aus Korrelationen entstehen, weil sie bestimmte Muster aufweisen, müssen nicht unbedingt kausal zusammenhängen oder, wenn bestimmte Eigenschaften in den Daten fehlen, können diese auch nicht zu einer idealen Lösung beitragen. Werden Daten in diesem Zusammenhang „als aufgezeichnete Fakten definiert, so sind Informationen die Menge von Mustern oder Erwartungen, die hinter diesen Daten stecken“ (Witten et al., 2017, S. xxiii). Die Datenqualität spielt eine wesentliche Rolle für unvoreingenommene Daten in KI-Modellen. Sie sind der Dreh- und Angelpunkt, wenn sie zu Mustererkennung mit ML für „Vorhersagen, Erklärungen oder Verständnishilfen“ (ebd.) im sozialen Kontext bspw. bei der Strafverfolgung, im Gesundheitssektor, in der Kreditvergabe usw. eingesetzt werden. Dabei spielen die ethischen Aspekte bei der Verwendung der Daten (insbesondere personenbezogene) in KI-Anwendungen eine große Rolle, die oftmals übersehen werden und zu Diskriminierung aufgrund von Verzerrungen in den Daten führen. Solche Verzerrungen werden *Bias* genannt, wenn antrainierte KI-Anwendungen aufgrund von geringer Datenqualität, die Voreingenommenheit beinhalten, und infolgedessen ungerechtfertigte Entscheidungen treffen. Prominente Beispiele sind der Einsatz von Software, die mit musterbasierte Prognosetechnik zukünftige Verbrechen auf der Grundlage historischer Verbrechensdaten vorhersagen (Perry et al. 2013, 1).

Da der Hauptfokus der vorliegenden Arbeit auf Daten liegt, werden im späteren Verlauf Bias-Effekte als Fallbeispiele betrachtet, die durch systematische Verzerrungseffekte in den zugrundliegenden Daten das Wissen und die Wahrnehmung (Baig 2010, S. 347) des ML einer KI-Anwendung beinträchtigen. Aber zunächst wird auf die systematische Diskriminierung eingegangen. Systematische Diskriminierung bezieht sich auf Strategien, Traditionen oder Verhaltensweisen, die Teil der Kultur oder Struktur einer Organisation sind (Mehrabi et al. 2021, S. 11). Mithilfe der systematischen Diskriminierung von bestimmten Individuen oder sozialen Gruppen einer Mehrheitsgesellschaft werden eben diese Strategien, Traditionen oder Verhaltensweisen aufrechterhalten (ebd.). Solche Praxen sind bspw. auf dem Arbeitsmarkt zu beobachten. Einer Studie (Veit et al. 2018, S. 35) zur Folge, bevorzugen Arbeitgebende in Deutschland eher Bewerbende, die ihnen kulturell ähnlich sind, wohingegen Bewerbende trotz Qualifikation Benachteiligung aufgrund ihrer Herkunftsländer, Religionszugehörigkeit und ihrem Phänotyp erfahren. In der Studie

wurde festgestellt, dass insbesondere Muslime und Menschen mit Schwarzem Phänotyp in Deutschland bei der Suche nach einem Arbeitsplatz diskriminiert werden. Solche diskriminierenden Praxen schlagen sich auch in Daten nieder, die sich in Trainingsdaten von KI-Anwendungen manifestieren und zu Bias-Effekten führen.

3.3.2 Diskriminierung durch voreingenommene Daten

Laut Hill Collins (2000) regeln die Unterdrückungsmechanismen einer Gesellschaft den Zugang zu Ressourcen, wenn soziale Gruppen aufgrund ihrer Zugehörigkeit über längere Zeit bevorzugt oder benachteiligt werden:

> "Oppression describes any unjust situation where, systematically and over a long period of time, one group denies another group access to the resources of society. *Race*, class, gender, sexuality, nation, age, and ethnicity among others constitute major forms of oppression in the United States." (Hill Collins 2000, S. 4)

Solche Unterdrückungsmechanismen werden durch KI-Anwendungen im folgenden Fallbespiel für Terrorverdacht eingesetzt. Laut einer Studie (2019) sammelt das Department of Homeland Security (DHS) in den USA personenbezogene Daten über die sozialen Medien und will diese noch ausweiten, um mit KI-Anwendungen die Sicherheitsrisiken von ausländischen Visaantragstellenden und amerikanischen Reisenden zu bewerten. Insbesondere wurde festgestellt, dass Menschen mit muslimischem Glauben besonders häufig gezielte Diskriminierung bei der Einreise oder Visaantragstellung durch KI-Bewertungen erfahren haben (Patel et al. 2019, S. 3).

Unterdrückungsmechanismen werden auch in Gesichtserkennungssystemen zur Überwachung eingesetzt. Sicherheitsbehörden in den USA setzen Gesichtserkennungssysteme mit KI-Anwendungen ein, um Verdächtige zu identifizieren (engl. face recognition). Bilder von verdächtigen Personen werden mit einer Datenbank abgeglichen, die vom Unternehmen ClearView erstellt wurden. Das Unternehmen hat laut Medienberichten (Brühl et al. 2020, o.S.) über drei Milliarden Aufnahmen von Gesichtern aus öffentlich zugänglichen Seiten in der Datensphäre, darunter Netzwerke wie Facebook, YouTube, Twitter und Instagram abgezogen und in interne Datenbanken gespeichert. Falls die KI-Anwendung eine Übereinstimmung findet, werden weitere Information an die Sicherheitsbehörden geliefert wie bspw. persönliche Daten oder die exakte Quellenangabe der Fotos.

Laut einer Onlinezeitung wurde nach einem fehlerhaften Abgleich mit einer Bilddatenbank von einem herangezoomten Ausschnitt eines Sicherheitsvideos, welches ein Überfall aufzeigt, der Afro-Amerikaner Robert Julian-Borchak Williams verhaftet. Die Polizeibehörde hatte sich auf den „Treffer" der KI-Anwendung ohne weitere Ermittlungen verlassen. Erst im Verhör sah Williams das Foto des gesuchten Straftäters und hielt es neben seinem Gesicht, und wies auf den eindeutigen Fehler hin:

> *"I picked it up and held it to my face and told him, 'I hope you don't think all Black people look alike"* (Allyn 2020, o.S.).

Im Folgenden Fallbeispiel geht es um automatische Gesichtsanalysesoftware, die bspw. in Smartphones zum Entsperren durch Gesichtserkennung vorhanden sind. Buolamwini (2018) deckt bei der automatischen Gesichtsanalysesoftware von verschiedenen Anbietenden Verzerrungen auf, die nicht ordnungsgemäß funktionieren. Buolamwini wird von einer Gesichtsanalysesoftware aufgrund ihres dunklen Hauttyps nicht erkannt, bis sie eine weiße Maske aufsetzt. Menschen mit dunklen Hauttypen sowie Frauen erkennt die KI-Anwendungen deutlich schlechter. Sie findet in ihrer Untersuchung heraus, dass bei hellen männlichen Gesichtern eine Fehlererkennungsrate bei weniger als 1 Prozent liegt, während bei hellen weiblichen Gesichtern die Fehlererkennungsrate auf etwa 8 Prozent steigt. Werden noch die phänotypischen Merkmale betrachtet, so liegt die Fehlererkennungsrate bei männlichen Gesichtern mit dunklem Hauttyp bei etwa 12 Prozent und bei weiblichen Gesichtern ist ein Anstieg von 35 Prozent zu verzeichnen (Buolamwini 2019, S. 2).

Warum *weiße* männliche Gesichter besser erkannt werden als alle anderen Gesichter, könnte bspw. daran liegen, dass 78 Prozent der KI-Fachkräfte weltweit männlich sind (World Economic Forum 2018, 28 f.). USA, Indien und Deutschland sind Länder mit dem größten Anteil an KI-Fachkräften, wobei Deutschland auch gleichzeitig zu den Ländern mit dem größten AI Gender Gap gehört (ebd.). Auch in Europa zeigt sich ein vergleichbares Bild. Nach eurostat (2022) pendeln die Anstellungsraten in IT-Berufen für männliche Personen im Jahr 2021 zwischen 70 bis 90 Prozent der westlichen Länder (eurostat 2022, o.S.). Wird der Blickwinkel um den Aspekt Ethnie erweitert, so fühlen sich laut einer Umfrage nur 16 Prozent der weiblich Befragten sowie Angehörige von ethnischen Minderheiten in Tech-Teams gut vertreten (Capgemini 2021, S. 2). Solche strukturellen Verzerrungen *nisten* sich als Bias-Effekte in die DNA einer KI-Anwendung ein, die dazu führen, dass die privilegierten Lebensrealitäten einer homogenen Gruppe von Menschen das Maß aller Werte und Normen bestimmen und die Ausschlusskriterien für die, ‚die in unerwünschter Weise anders' sind (Scherr 2010, S. 43).

Auch im Bewerbungsverfahren spiegelt sich die gesellschaftliche Realität wider. Einem online Zeitungsartikel zu Folge (Dastin 2018) hat Amazon, KI-Anwendungen mit ML eingesetzt, um aus eingehenden Bewerbungen die besten Kandidat*innen für eine Stellenausschreibung automatisiert herauszufiltern. Das Wissen der antrainierten KI-Anwendung bestand aus Lebensläufen der letzten zehn Jahren, die erfolgreich eingestellt wurden. Nach einem Jahr Einsatz fiel dem Entwickler*innenteam auf, dass das Sortierverfahren des ML sich bei technischen Stellenausschreibungen nicht geschlechtsneutral verhielt, weil die erfolgreich eingestellten Lebensläufe der letzten

zehn Jahre männlich waren. Die Daten wurden daraufhin modifiziert und Geschlechtsangaben anonymisiert. Trotz dieser Anonymisierung hat das ML Methoden gefunden, Frauen zu diskriminieren, indem Lebensläufe aussortiert wurden, die geschlechtsspezifische Hinweise enthielten wie bspw. ‚Frauen Schachclub' (Dastin 2018, o.S.).

Dieser Hinweis, der auf das Geschlecht Rückschlüsse ziehen lässt, hat die KI-Anwendung trotz Anonymisierung erkannt, weil sie im Sortierverfahren nach Proxies gesucht hat, die mit dem Geschlecht korrelieren (Matzner 2022, o.S.).

China ist beim Thema Überwachung eines der Länder, welches nahezu in allen Lebensbereichen KI-Anwendungen einsetzt. KI-Anwendungen sammeln, speichern und werten Daten von Bürger*innen aus, um die Menschen systemkonform zu erziehen. Solch ein soziales Kreditsystem (engl. Social Credit) ist ein verhaltensbelohnendes oder -sanktionierendes Punktevergabesystem, welches zum Überwachen durch datengetriebene KI-Anwendungen eingesetzt wird. Dieses Überwachungssystem findet seine Wurzeln bereits in den 1990er Jahren, als versucht wurde, persönliche Bank- und Kreditbewertungssysteme zu entwickeln, um insbesondere die Kreditvergabe in ländlichen Gebieten zu erleichtern, in denen Einzelpersonen und kleine Unternehmen oft keine dokumentierte finanzielle Vergangenheit hatten (Drinhausen et al. 2021, S. 4). 2007 fand eine Ausarbeitung des Sozialen Kreditsystems durch staatliche Stellen statt, wobei sich ab 2011 die Zielsetzung von einem Finanzkredit-Rating-System zu einem Allzweckmittel sozialpolitischer Maßnahmen gewandelt hat (ebd.).

3.3.3 Fairness als Kompetenz für KI-Anwendungen

Wie in den vorangegangenen Kapiteln aufgezeigt, ist gerechtes Verhalten einer KI-Anwendung eine Schlüsselkompetenz, wenn sie als soziotechnische Technologie in einer Gesellschaft eingebettet wird. Die EU will ein KI-Ökosystem aufbauen, welches die Grundlagen für vertrauenswürdige KI-Anwendungen (HEG-KI 2019) schaffen will, weil KI-Anwendungen aufgrund von schlechten Datengrundlagen diskriminieren können, wie die Fallbeispiele weiter oben aufgezeigt haben. Zugleich stellt die EU jedoch in einer Umfrage fest, dass EU-weit Menschen aus Nordafrika, aus der Subsahara und Rom*nja am meisten aufgrund ihres Aussehens diskriminiert werden, während Menschen mit Migrationsgeschichte und deren Nachkommen aus Nordafrika und der Türkei häufiger Diskriminierung aufgrund ihres Namens erfahren (FRA 2017, 13 f.). Einerseits will die EU vertrauenswürdige KI-Anwendungen in einem KI-Ökosystem ausbauen, andererseits zeigen die realen Zustände einer Gesellschaft, dass weder sie diskriminierungsfrei noch unvoreingenommen ist. Es wird jedoch die Anforderung an KI-Anwendungen gestellt, die vorherrschende Diskriminierung in einer Gesellschaft nicht zu übernehmen. Die Trainingsdaten, die einer KI-Anwendung mit ML zur Verfügung

gestellt werden, die die gesellschaftlichen Vorurteile mit patriarchalen Machtstrukturen, neoliberalem Kapitalismus oder das Erbe des Kolonialismus (► vgl. Kap. 3.2.3) widerspiegeln, sollen jedoch die Ansprüche der Fairness erfüllen, um gerechte Entscheidungen zu treffen.

Was also bedeutet Fairness, die eine Schlüsselkompetenz für vertrauenswürdige KI-Anwendungen ist? Wenn die Wortbedeutung des Begriffs Fairness betrachtet wird, so kommt der Begriff Fairness aus dem Englischen und bedeutet einwandfreies, anständiges Verhalten (DWDS 2022, o.S.). Demnach sollen KI-Anwendungen mit ML sich einwandfrei und anständig Verhalten, wenn sie Entscheidungen, Prognosen oder Bewertungen über Menschen und deren Verhalten treffen. Nach Dimitriou et al. (2015) ist Fairness „ein zentrales philosophisches Konzept, insbesondere in der Ethik- und der Gerechtigkeitstheorie“, die sich mit der Fragestellung des Miteinanders trotz Differenzen nach ethischen Maßstäben beschäftigen, wobei „die Ethik (…) zumeist einen Fokus auf das Handeln der Personen abzielt, [während] (…) die Gerechtigkeitstheorie [sich] vor allem mit Fragen der Einrichtung, der Konstitution der Gesellschaft und der Verteilung wichtiger Güter durch den Staat [befasst]“ (Dimitriou et al. 2015, S. 15). Daher ist nach den Publizierenden „ein gesellschaftlicher Zustand dann gerecht, wenn er fair [ist]“ (ebd.). Auf den unfairen gesellschaftlichen Zustand weist der Feminismus seit seiner Entstehung hin, indem er die Demokratisierungsdefizite einer gesellschaftlichen Konstitution identifiziert und Widerstandsformen gegen die Unterdrückung entwickelt. Das Grundkonzept der feministischen Strömungen fordert Gerechtigkeit, die für alle Menschen gilt, so wie das Konzept der Fairness. Somit ist der Feminismus mit intersektionaler Perspektive für die Fairness einer Mehrheitsgesellschaft, die unvoreingenommen und diskriminierungsfrei gegenüber sozialen Gruppen oder Minderheiten ist. Womit die Fairness für feministische KI-Anwendungen als Schlüsselkompetenz gilt. Hiermit ist nun die theoretische Rahmung für die vorliegende Thesis vollständig. Aus diesem theoretischen Bezugsrahmen wird im Folgenden das Framework gebildet, welches zum einen zur deduktiven Kategorienbildung dient und zum anderen zur Strukturierung der Erhebungsergebnisse (Bortz et al. 2016, S. 330)

3.4 Framework zur Kategorienbildung

In diesem Kapitel wird das Framework für die systematische Literaturanalyse sowie für die Forschungsfrage gesetzt. In den vorangegangenen Kapiteln wurde festgestellt, dass der Begriff *Künstliche Intelligenz* keine allgemeingültige Definition hat. Während sozialwissenschaftliche bzw. ethische Diskussionen den Begriff KI verwenden, ist im Ingenieur*inwesen die Rede vom *Maschinellen Lernen*. Der Fokus liegt auf ethischen Aspekten einer datengetriebenen KI-Anwendung. Technische Publikationen,

die technische Lösungen thematisieren, werden für den **ersten Eckpunkt** exkludiert. Denn in der Phase der Recherche und Vorbereitung dieser Thesis wurde festgestellt, dass für Verzerrungen oder Bias in Daten keine Lösungen existieren. Insbesondere macht der Kriterien-Katalog AIC4 des Bundesamts für Sicherheit in der Informationstechnik (BSI) dies deutlich. In diesem Katalog wird die Aussage getroffen, dass Bias in KI-Anwendungen oft mit moralischen oder ethischen Fragen wie der fairen und gerechten Behandlung von Individuen oder sozialen Gruppen verbunden seien:

> "The topic of bias in AI applications is often linked to moral or ethical questions like the fair treatment of individuals or groups. The BSI does not make any statements regarding ethical questions" (BSI 2021a, S. 42).

Daher ist **der zweite Eckpunkt** für das Framework zur Identifizierung von Literatur mit KI gesetzt. Zudem sind hohe Investitionen sowie Patente ab 2019 festzustellen. Daher ist **der dritte Eckpunkt** der Zeitrahmen für die Literaturrecherche, der von 2019 bis Mai 2022 gesetzt wird.

Der Einsatz von KI-Anwendungen wird im Hinblick auf die Ökonomie thematisiert. Wohingegen zivilgesellschaftliche Diskurse über gemeinwohltätige Einsätze von KI-Anwendungen kaum thematisiert werden. Der Stand der heutigen KI-Anwendungen mit ML entspricht der einer schwachen KI und ist nichts anderes als Software, die zunächst mit allgemeinen Algorithmen (je nach Aufgabenstellung, ob Klassifikation oder Regression) modelliert wird und mit Trainingsdaten ein Regelwerk für den zweiten Algorithmus erstellt, um zukünftig mit dem gelernten Regelwerk über neue Daten Entscheidungen oder Vorhersagen zu treffen. Wenn in den Programmcode beim Modellieren einer KI-Anwendung keine explizite Diskriminierung eingebaut wird und der Programmcode Softwarepackages bezieht, die aktuell und transparent sind, dann sollten die algorithmischen Entscheidungssysteme (Algorithmic decision making systems, ADM Systems) weitestgehend fehlerfrei funktionieren. Womit hier **der vierte Eckpunkt** gesetzt wird, der da wäre, dass Algorithmen aus der Literaturrecherche exkludiert, aber für **den fünften Eckpunkt** die Trainingsdaten inkludiert werden, die als Datengrundlage für die Entwicklung des Regelwerks dienen. In der Sphäre des Internets werden digitale Daten von und über Nutzende produziert und sind von besonderem Interesse für verschiedene Akteur*innen der Gesellschaft, weil sich aus diesen Daten neue Erkenntnisse und Wissen ableiten lassen. Daher folgt nun **der sechste Eckpunkt**, die die Informationsverarbeitung im Big Data Verfahren darstellt, mit den Punkten der Datenerhebung und Datensammlung von verschiedenen Datenstrukturen, jedoch nicht der Analyse und Auswertung im Data Mining Verfahren, da diese durch den Algorithmus durchgeführt wird und der wurde weiter oben exkludiert. **Der siebte Eckpunkt** wird mit Datenqualität gesetzt, die die Grundlage für die korrekte Arbeitsweise einer antrainierten KI-Anwendung bildet. Der Umgang mit Daten von hoher Qualität geht mit ethischen

Aspekten einher, die der Feminismus bietet. Daten, die von verschiedenen Akteur*innen der Gesellschaft erhoben und gesammelt werden, stellen den strukturellen Rahmen dar, der den Grad der Bias-Effekte in KI-Anwendungen bestimmt. Der feministische Ansatz weist auf die Abwesenheit von Selbstbestimmung, Freiheit und Gleichheit hin und verdeutlicht die Demokratisierungsdefizite einer Gesellschaft, die Gerechtigkeit verspricht. Womit nun der **achte Eckpunkt** hier gesetzt wird, um die pre-existing Bias für die systematische Literaturanalyse einzurahmen, die in die zugrundeliegenden Trainingsdaten einfließen und die Werte einer Gesellschaft widerspiegeln. Das ML in KI-Anwendungen übernimmt die Werte und entwickelt entsprechend das dazugehörige Regelwerk, was zu technischen Bias-Effekten führt. Daher werden Bias-Effekte, die technischer Natur sind, nicht berücksichtig, denn mathematische Konstrukte brauchen Formeln und Zahlen, um Funktionen korrekt zu berechnen und eindeutige Eingabedaten. Nach der Umwandlung des Sprichworts „Fine Feathers Make Fine Birds“ würde es dann bedeuten: Fine Dataset Make Fine AI Application“. Daher liegt der Fokus auf der Ursache und Behebung von pre-existing Bias in den Trainingsdaten. Aus diesem Framework ergeben sich vier Kategorien:

Datengrundlage, Datenerhebung, Datensammlung und Datenqualität

Diese deduktiven Kategorien, die theoriegeleitet gebildet wurden, dienen zur Strukturierung der Erhebungsergebnisse. Mit diesen Erkenntnissen wird als nächstes auf die Methodik der systematischen Literaturanalyse eingegangen.

4. Methodik

Dieses Kapitel beschreibt die Vorgehensweise der verwendeten Methoden sowie deren Umsetzung. Zunächst wird die Forschungsmethode erläutert. Im Anschluss daran wird die Forschungsfrage mit einem Schema formuliert und die Bildung der Fragestellung aufgezeigt. Ferner wird der detaillierte Forschungsablauf beschrieben, der die Ein- und Ausschlusskriterien beinhaltet sowie den Kriterienkatalog zur Qualitäts- und Ausschlussbewertung. Darüber hinaus wird die Suchstrategie für die datenbankgestützte Recherche dargestellt sowie die ausgewählten Datenbanken vorgestellt. Dann folgen die Ergebnisse der Literaturauswahl und abschließend wird die Vorgehensweise für die Datenextraktion sowie das Auswertungsvorgehen dargelegt.

4.1 Forschungsmethode

In der vorliegenden Arbeit wird mithilfe der systematischen Literaturanalyse (engl. Systematic Literature Reviews, SLR), angelehnt an Wetterich et al. (2021) eine „datenbankgestützte und Algorithmus geleitete Recherche mit der daran anschließenden [qualitativen] Analyse“ (ebd., S. 10) durchgeführt. Laut den Publizierenden Wetterich et al. (2021, S. 20) kommt diese Methode aus der evidenzbasierten Medizin. Erst im Jahr 2006 fand die Methode des SLRs den Zugang zu den Sozialwissenschaften (ebd.). Der methodische Ansatz der SLR bietet nicht nur eine systematisierte Literaturanalyse, sondern auch einen transparenten analytischen Rahmen und beruht auf einem nachvollziehbaren Algorithmus (ebd.). Dies ist deshalb wichtig, weil im Jahr 2021 insgesamt 334 497 KI-Publikationen in englischer Sprache weltweit veröffentlicht wurden (Zhang et al. 2022, S. 17). Um aus der Masse der Veröffentlichungen die relevanten Publikationen zu identifizieren, ermöglicht die Methode der SLR „größtmögliche Transparenz und Objektivität in der Literaturarbeit“ (Wetterich et al. 2021, S. 10). Bevor jedoch die SLR beginnt, wird im Folgenden angelehnt an Wetterich et al. (2021, S. 27) die Forschungsfrage nach dem PICO-Schema hergeleitet.

4.2 Forschungsfrage

Diskriminierende KI-Anwendungen haben in der Vergangenheit aufgezeigt, dass sie aufgrund von verzerrter Datengrundlage voreingenommen über Menschen entscheiden oder Vorhersagen über ihr Verhalten treffen (► vgl. hierzu Kap. 3.3.2). Oftmals werden die lernenden Algorithmen in KI-Anwendungen für Diskriminierungen oder Verzerrungen verantwortlich gemacht, die aufgrund der Datengrundlage fehlerhaftes Verhalten aufzeigen, weil entweder nicht alles Wissen der Welt in den Trainingsdaten abgebildet werden können oder reale Zustände der Ungerechtigkeit in

der Welt sich in der Datengrundlage widerspiegeln (► vgl. bspw. Matzner 2022, Mittelstadt 2016, Zweig 2019). Daher liegt der Fokus auf personenbezogenen Daten. In der folgenden Tabelle wird, basierend auf dem PICO-Schema (**P**opulation, **I**ntervention, **C**omparison und **O**utcome), die Forschungsfrage hergeleitet und ausformuliert:

Tabelle 4.1: Forschungsfrage nach PICO-Schema

Entwicklung der Forschungsfrage nach PICO-Schema	
***P**opulation: An welchen Daten liegt das Erkenntnisinteresse?*	Daten, die personenbezogen sind und mit denen KI-Anwendungen mit ML trainiert werden.
***I**ntervention: An welcher Intervention liegt das Erkenntnisinteresse?*	Ursache von pre-existing Bias in Daten.
***C**omparison: Mit welcher Intervention wird verglichen?*	Mit KI-Anwendungen, die im Kap. 3.3.2 dargestellt wurden.
***O**utcomes: Welche spezifische Auswirkung der Intervention steht im Fokus?*	Behebung von pre-existing Bias, die in Daten enthalten sind.
Forschungsfragen für die systematische Literaturanalyse	
Ursache	*Lösung*
► Was sind die Ursachen von pre-existing Bias in Daten?	► Was ist in der wissenschaftlichen Literatur zur Behebung von pre-existing Bias in Daten bekannt? ► Welche Konzepte oder Modelle werden in den Publikationen vorgeschlagen, die pre-existing Bias in Daten beheben?

Um die Behebung von pre-existing Bias zu verstehen, wird die mögliche Erklärung für die Ursache betrachtet, um im nächsten Schritt die Lösungen zu analysieren. Das Forschungsziel wurde bereits im Kapitel 1.2 geschildert. Mit der Forschungsfrage soll herausgefunden werden, auf welche Weise sich die Literatur zum Themenkomplex bezieht (Wetterich et al. 2021, S. 28). Hierzu wird nun durch den Forschungsablauf die genaue Vorgehensweise dokumentiert und dargestellt.

4.3 Forschungsablauf

Der Forschungsablauf dient der Nachvollziehbarkeit der systematischen Literaturrecherche und -analyse, dessen Durchführung und methodisches Vorgehen protokolliert wird, um größtmögliche Transparenz sicherzustellen (Wetterich et al. 2021, S. 29). Im ersten Schritt werden Einschlusskriterien definiert, die der Auswahl der Publikationen dienen. Nach diesem Vorgang erfolgt die Datenanalyse, die mit einem Kriterienkatalog die Qualität von den ausgewählten Publikationen bewertet, die ausgeschlossen bzw. eingeschlossen werden. Mit den übrig gebliebenen Publikationen,

die für die vorliegende Arbeit relevant sind, erfolgt die Datenextraktion mit qualitativer Inhaltsanalyse.

4.3.1 Ein- und Ausschlusskriterien

Die Einschlusskriterien für die Auswahl der Publikationen beinhalten als erstes die Publikationsart. Es sollen wissenschaftliche Publikationen sowie ‚graue Literatur' eingeschlossen werden, die bspw. von verschiedenen Organisationen, Institution usw. veröffentlicht wurden. Ferner werden nur Publikationen berücksichtigt, die in der englischen Sprache im Zeitrahmen von 2019 bis 2022 veröffentlicht wurden. Der Suchalgorithmus wird im Abstract angewandt.

Tabelle 4.2: Einschlusskriterien der Publikationen

Schlüsseldaten	Kriterien
Zeit	Von 2019 bis 2022
Publikationsart	Wissenschaftliche Publikationen
Anwendung des Suchalgorithmus	Abstract, Title und Keywords
Sektor	Computer Science, Social Science, Philosophy
Sprache	Englisch
Outcome	Behebung von pre-existing Bias, die in Daten enthalten sind.
Kontext	Daten in KI-Anwendungen mit ML.

Im nächsten Schritt wird für die Datenanalyse ein Kriterienkatalog für die Qualitätsbewertung erstellt, die zur Aussortierung der ausgewählten Publikationen herangezogen wird. Der Kriterienkatalog wird im Abstract der Publikationen angewandt. Hierbei erstreckt sich der Kriterienkatalog auf zwei Dimensionen. Zum einen wird die Intervention begutachtet, die die Ursachen von pre-existing Bias in den Daten thematisiert. Zum anderen wird das Outcome bewertet, wenn Konzepte oder Modelle in der Abhandlung der Publikation vorhanden sind. Die Punktevergabe für die Beantwortung der Fragen im Kriterienkatalog erfolgt mit **J**A=1, **N**EIN=0 und **T**EILWEISE=0,5, wobei das Mindestergebnis der Bewertung auf 1,5 gesetzt wird, um in die weitere Verarbeitung der SLR eingeschlossen zu werden.

Tabelle 4.3: Kriterienkatalog zur Qualitäts- und Ausschlussbewertung

Dimension	Kriterienkatalog	Ranking
Intervention	Werden Ursachen für pre-existing Bias in Daten thematisiert?	J, N, T
Outcome	Werden Lösungen für pre-existing Bias in Daten thematisiert?	J, N, T

4.3.2 Auswahl von Datenbanken

In der folgenden Tabelle sind drei Online-Datenbanken aufgelistet, mit denen die Literaturrecherche durchgeführt wird. Die Publizierenden Wetterich et al. (2021, S. 46) empfehlen, sich auf Online-Datenbanken zu konzentrieren. Zum einen sind in Online-Datenbanken „Informationen im Normalfall aktuell“ und zum anderen „schnell zugänglich“ (ebd.). Die Anwendung des Suchalgorithmus wird in den Abstracts, Title und Schlagwörtern durchgeführt. Die Online-Datenbanken, die im Folgenden aufgeführt werden, bieten bibliografische Einträge an, da sie „eine organisierte digitale Sammlung von Verweisen auf veröffentlichte Literatur, einschließlich Zeitschriften- und Zeitungsartikeln, Konferenzberichten, Berichten, Regierungs- und Rechtspublikationen, Patenten, Büchern usw.“ enthalten (ebd.). Die Auswahl der Datenbanken erfolgte aufgrund ihrer Reichweite und ihrer Nutzung von Webcrawlern. Als Software in Suchmaschinen durchsucht ein Webcrawler automatisch das gesamte Internet nach den gesuchten Suchbegriffen. Die Datenbanken von ACM Digital Library, BASE (Bielefeld Academic Search Engine) Semantic Scholar greifen bspw. auf alle Dokumente von ArXiv.org (Cornell University Library) oder ScienceDirect oder Elservier usw. zu, weshalb auf diese verzichtet wurde, um größtmögliche Auswahl an Publikationen zu erhalten.

Tabelle 4.4: Auswahl der Datenbanken

Datenbanken	Beschreibung
ACM Digital Library (https://www.acm.org/publications/digital-library)	The ACM Digital Library (DL) is the world’s most comprehensive database of full-text articles and bibliographic literature covering computing and information technology. This renowned repository includes the complete collection of ACM publications plus an extended bibliographic database of core works in computing from scholarly publishers.
BASE (Bielefeld Academic Search Engine) (https://www.base-search.net/)	BASE (Bielefeld Academic Search Engine) ist eine der weltweit größten Suchmaschinen für wissenschaftliche Web-Dokumente. Der Index umfasst über 240 Millionen Dokumente von über 8.000 Datenlieferanten. Bei etwa 60% der in BASE indexierten Dokumente sind die Volltexte frei zugänglich (Open Access). Betreiberin der Suchmaschine BASE ist die Universitätsbibliothek Bielefeld.
Semantic Scholar (https://www.semanticscholar.org/)	Semantic Scholar is a free, AI-powered research tool for scientific literature, based at the Allen Institute for AI. The tool index over 200 million academic papers sourced from publisher partnerships, data providers, and web crawls.

4.3.3 Suchstrategie

Für die Suche in den Datenbanken werden verschiedene Boolesche Operatoren zur datenbankgestützten und Algorithmus geleiteten Recherche verwendet. Die Suchoperatoren, welche in der folgenden Tabelle aufgelistet sind, ermöglichen die Kombination von mehreren Suchbegriffen sowie die Anpassung der Recherche (Neumann 2018, S. 4), um konkrete Suchergebnisse zu erhalten. Wie in den vorangegangenen Kapiteln aufgezeigt, gibt es eine Vielzahl an Publikationen rund um KI. Um nur die relevanten Publikationen zu filtern, die für die Beantwortung der Forschungsfrage herangezogen werden sollen, sind auch mehr als nur die gängigen Suchoperatoren wie AND, OR, NOT notwendig. Daher werden in der folgenden Tabelle die Suchoperatoren und ihre Funktionen beschrieben:

Tabelle 4.5: Suchoperatoren

Operatoren	Beschreibung
+ AND	Um gezielter zu suchen, werden mehrere Begriffe miteinander verbunden, die alle im Suchergebnis vorkommen sollen. Z.B. “data+ethic”
- NOT	Wird verwendet, wenn Suchbegriffe ausgeschlossen werden sollen. z.B. “data -Algorithm”. Es werden nur Suchergebnisse ausgegeben, die den Begriff data haben, aber ohne Algorithm
| OR	Mit dem Pipe Symbol werden Begriffe separiert z.B. “data + ethic | unbiased” (entweder das eine oder das andere),
„…“ exakte Phrase	Es wird mit der exakten Phrase gesucht, z.B. “ machine learning unbiased data“ → „AI machine learning unbiased data“ würde nicht angezeigt werden, da ‚AI‘ in der Phrase nicht enthalten ist.
* Wildcard	Mit der Wildcard Suche werden Begriffe ersetzet, die nicht bekannt sind, z.B. “data*” findet Datensätze mit datadriven, dataset, datamanagement usw.
? *einzelnes Zeichen*	Mit dem Fragezeichen ? als Platzhalter wird ein einzelnes unbekanntes Zeichen angegeben, da der Begriff in unterschiedlichen Schreibweisen existiert, hier durch ein Leerzeichen oder Bindestrich z.B. pre?existing findet das Zeichen, das zwischen den zwei Wörtern in der Datenbank existiert.

In der folgenden Tabelle wird nun der Suchalgorithmus auf der Grundlage der Forschungsfrage entwickelt. Ausgehend von der Hauptfrage sind in der linken Spalte der Tabelle die Themen enthalten, für die die Operatoren eingesetzt werden. In der letzten Spalte werden die Suchterme zum Suchalgorithmus zusammengesetzt.

Tabelle 4.6: Suchterme mit Operatoren

Hauptfrage ► Was ist in der wissenschaftlichen Literatur zur Behebung und Ursache von pre-existing Bias in Daten bekannt?	
Themen	Suchterm mit Operatoren
pre-existing Bias	„pre?existing bias“ oder „pre-existing bias“
historische Daten	„historical data“
Unvoreingenommene diskriminierungsfreie Daten	fair* | discrimina* | ethic*
Trainingsdaten Datensammlung Datenerhebung	data* → damit würden alle Publikationen als Suchergebnis ausgegeben werden, die die Begriffe ‚data set‘ ‚data collection‘ ‚data handling‘ usw. enthalten.
Vertrauenswürdige, erklärbare, verantwortungsbewusste, interpretierbare, nachvollziehbare soziotechnisch KI	trustwort* | explain* | responsi* | interpreta* | comprehen* + AI | “Artificial Intelligence”

Diese Suchterme werden in den ausgesuchten Datenbanken zur Literaturrecherche angewandt. Jede Datenbank hat eine eigene innere Logik und verhält sich unterschiedlich. Während die Datenbank von *ACM Digital Library* mehrfache Operatoren zulässt, um die Suche feingliedriger durchzuführen, geben die Datenbanken von *Semantic Scholar* und Base ‚null‘ Suchergebnisse aus. Daher werden die Suchterme entsprechend der Datenbanklogik angepasst.

Tabelle 4.7: Suchalgorithmen

Datenbanken	Suchalgorithmus
ACM Digital Library	„pre?existing bias“ + „historical data“ + fair* | discrimina* | ethic* + data* + AI | “Artificial Intelligence” + trustwort* | explain* | responsi* | interpreta* | comprehen* ►*Algorithmus funktioniert zwar vollständig aber die Datenbank unterteilt trotz Filtersetzung des Zeitrahmens Publikationen bzw. Konferenzen nach Jahren und erlaubt keine Mehrfachauswahl. Suchergebnisse mit Literaturangaben können alle auf einmal ausgewählt und in eine BibTeX Datei exportiert werden (s. Anhang 1).*
BASE (Bielefeld Academic Search Engine)	pre-existing OR bias* OR histor* OR data* OR fair* OR trustwort* OR comprehen* OR interpreta* OR respon* OR discrimina* AND artificial intelligence OR AI ►*Werden die Begriffe ethic* OR explain* zusätzlich zum Suchalgorithmus in den Suchalgorithmus eingeschlossen, gibt es keinen Treffer. Filter für den Zeitrahmen (von bis) kann nicht gesetzt werden. Publikationen können nur nach Erscheinungsjahren gefiltert werden. Zehn Sucherergebnisse pro Seite können angewählt werden und als BibTex heruntergeladen werden (s. Anhang 2).*
Semantic Scholar	pre-existing bias* histor* data* artificial intelligence fair* ►*Schwerfällig im Umgang und gibt beim obigen ‘langen‘ Algorithmus keine Treffer im Suchergebnis zurück. Deshalb musste der Algorithmus gekürzt werden. Operatoren wie | OR werden zu + umgewandelt und schränken die Suchergebnisse ein. Fragezeichen wird im Suchfeld nicht angenommen. Suchergebnisse mit Literaturangaben können alle auf einmal ausgewählt und in eine BibTeX Datei exportiert werden (s. Anhang 3).*

4.4 Ergebnisse zur Auswahl der Publikationen in den Datenbanken

In diesem Kapitel werden die Auswahl der Publikationen in den drei Datenbanken beschrieben.

4.4.1 ACM Digital Library

Die Auswahl der ACM (Association for Computing Machinery) Digital Library ist deshalb getroffen, weil sie eine Datenbank der wissenschaftlichen Publikationen zu Themen rund um Informatik beherbergt. Die Literaturreche wurde am 13.06.2022 durchgeführt. Die Suche ergab, mit einem Zeitfilter von 2019 bis 2022, einen Treffer von 7134. Da diese Datenbank unterschiedliche Dateiformate ausgibt, wurde ein Filter für Dateiformate gesetzt, sodass die Suchtreffer ausschließlich PDF-Dateien enthalten. Damit fiel die Trefferzahl auf 6988. Ein weiterer Filter wurde angewandt, um wissenschaftliche Artikel zu erhalten, womit sich die Trefferzahl auf 5204 beläuft. Mit *Special Interest Group on Artificial Intelligence* (SIGAI) wurde eine weitere Filterung angewandt, um die Suchergebnisse weiter einzugrenzen. Damit ergaben sich 302

Treffer. Mit dem Themenfeld *Association For Computing Machinery* wurde die Suche nochmals auf 245 Treffer eingegrenzt. Die Datenbank bietet ab diesem Schritt nur noch einzelne Oberthemen nach Jahren, die gefiltert werden können, obwohl im ersten Schritt der Zeitrahmen gesetzt war. Deshalb wurde die Filterung einzeln auf das Themenfeld ‚AI, Ethics, And Society (AIES)' angewandt, womit für AIES'21 - 74 Treffer, AIES'20 - 43 Treffer und AIES'19 – 39 Treffer ergab. Da es für AIES keine Publikationen für das Jahr 2022 gab, wurde ein weiteres Themenfeld ‚Transactions On Knowledge Discovery From Data (TKKD)' mit 34 Treffern hinzugezogen. Publikationen, die bspw. andere Themenfelder wie die Gesetzgebung thematisierten, wurden nicht berücksichtigt, womit aus der ACM Digital Library genau 190 Publikationen übriggeblieben sind. Im Anhang 1 befinden sich in der Tabelle I die Selektionsvorgehensweise der Publikationen mit den dazugehörigen Screenshots sowie die Literaturangaben, die als BibTeX heruntergeladen wurden.

4.4.2 BASE (Bielefeld Academic Search Engine)

Die Suchmaschine BASE (Bielefeld Academic Search Engine) wird von der Universitätsbibliothek Bielefeld betrieben, die von 8000 Datenlieferanten wissenschaftliche Web-Dokumente bezieht. Die Literaturrecherche fand am 15.06.2022 statt. Zunächst wurde der Suchalgorithmus ohne Filter in das Suchfeld eingegeben, da die Suchmaschine keine Filterung eines Zeitrahmens zulässt, womit eine Treffermenge von 278 Publikationen erzielt wurde. Daher wurde für jedes Erscheinungsjahr einzeln die Suche durchgeführt, womit sich von 2019 bis 2022 insgesamt 112 Publikationen ergaben. Eine weitere Filterung nach Dokumentenart ‚Artikel' hat 54 Treffer ergeben. Im letzten Schritt wurden nur noch solche Publikationen ausgewählt, die in englischer Sprache veröffentlich worden sind, womit für die Erscheinungsjahre 2019 bis 2022 insgesamt 40 Publikationen übriggeblieben sind. Die vollständigen Literaturangaben für diese Suche sind im Anhang 2 aufgelistet sowie die Selektionsvorgehensweise der Publikationen mit den dazugehörigen Screenshots in der Tabelle II.

4.4.3 Semantic Scholar

Semantic Scholar ist eine KI-gesteuerte Suchmaschine für akademische Publikationen, die am Allen Institute for AI entwickelt wurde. Die Literaturrecherche fand am 14.06.2022 statt. Obwohl der Suchalgorithmus für diese Datenbank sehr allgemein gehalten wurde, hat die Datenbank nach der Filterung des Zeitrahmens und Dateiformat lediglich 114 Treffer erzielt. Im nächsten Schritt wurde nach diesen Themenbereichen (Field of Study) gefiltert: Computer Science, Philosophy und Sociology. Dies ergab eine Trefferzahl von 84. Darauffolgend wurde eine weitere Filterung von Publication Type angewandt. Ausgewählt wurden Journal Article, Review und Conference Typen, was 68 Treffer ergab. Im letzten Schritt wurde im Filterbereich „Journals & Conferences"

folgende Themenbereiche ausgewählt: ArXiv (Dokumentenserver), AIES (AI, Ethics, and Society) und KDD (Knowledge Discovery and Data Mining). Insgesamt ergaben sich mit der Suchmaschine von Semantic Scholar 28 Publikationen, die für SLR verwendet werden. Im Anhang 3 befinden sich die dazugehörigen Literaturangaben sowie in der Tabelle III die Auswahlschritte, die mit Screenshots dokumentiert sind.

4.5 Screening Prozess

Zusammenfassend wurden insgesamt 258 Publikationen durch die systematische Einschränkung der Suchergebnisse identifiziert. In Abbildung 4.1 wird der Screening Prozess graphisch dargestellt. Im nächsten Schritt wurden alle Duplikate entfernt, womit 239 Publikationen übriggeblieben sind. Der Auswahlprozess von Publikationen besteht aus zwei Phasen (Wetterich et al. 2021, S. 59). Die erste Phase beinhaltet die Sichtung und Sortierung von Titeln und Abstracts aller Publikationen, die den vorgegebenen Inklusionskriterien entsprechen (ebd.). Diese Inklusionskriterien sind die Interventions und Outcomes, die entsprechend ihren Eignungskriterien mit Punkten vergeben wurden (► vgl. Tabelle 3 im Kap. 5.3.1). Hiernach wurden alle Publikationen ausgeschlossen, die eine Punktevergabe unter 1,5 Punkten erhielten. Damit sind insgesamt 173 Publikationen aus der SLR ausgeschlossen. In der zweiten Phase wurde der Volltext aller übriggebliebenen 66 Publikationen jedes identifizierten Artikels analysiert und nach den Ein- und Ausschlusskriterien beurteilt. Publikationen, die im Abstract Hinweise auf mögliche Antworten der Forschungsfragen gaben, jedoch im Volltext nicht, wurden von der SLR ausgeschlossen, womit 38 Publikationen ausgeschlossen wurden. Im Anhang 4 wird in Tabelle IV der Ausschlussgrund mit Kommentaren dargestellt. Schlussendlich sind 28 Publikationen in die SLR aufgenommen wurden, die in der Tabelle V im Anhang 5 mit ihrer Punktebewertung dargestellt werden. Darüber hinaus werden die

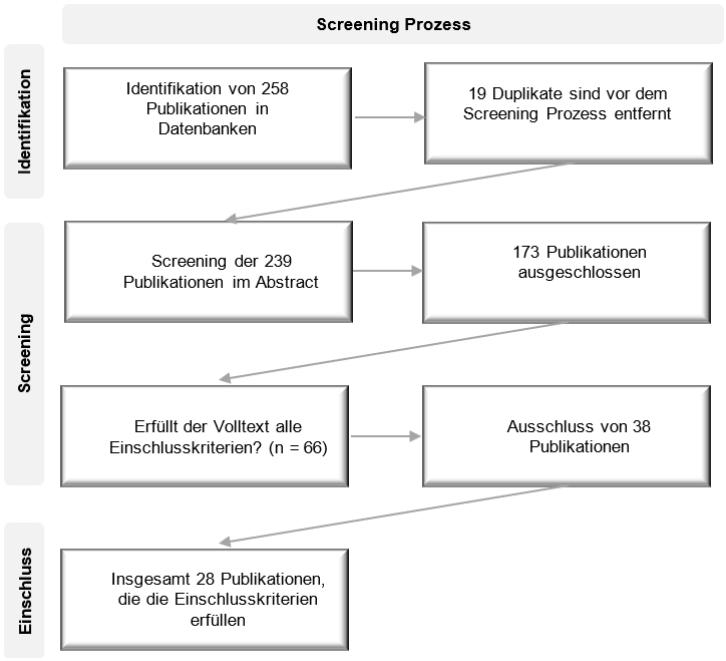


Abbildung 4.1: Screening Prozess (in Anlehnung an: Page et al. 2021)

ausgewählten Publikationen im Anhang 6 in der Tabelle VI mit Zielen und Ergebnissen ihrer Untersuchungen kurz vorgestellt.

4.6 Datenextraktion und Auswertungsvorgehen

Nach dem Screening Prozess erfolgt die Datenextraktion. Mit den Methoden der qualitativen Inhaltsanalyse nach Mayring (2015) werden die erhobenen Daten mithilfe der Software MAXQDA ausgewertet. Das Datenmaterial besteht aus 28 Publikationen, die einer qualitativen Inhaltsanalyse unterzogen werden, in dem sie strukturierend und analysierend auf relevante Stellen für die Beantwortung der Forschungsfragen kodiert werden. Dabei wird die Methode nach Mayring (2015, S. 52) angewandt, um systematisch bei der Inhaltanalyse vorzugehen, die wissenschaftliche Reproduzierbarkeit sicherstellt. Das theoriegeleitete Framework wurde im Kapitel 3.4 für das Kategoriensystem gesetzt, welches das Werkzeug für die Kodierung darstellt. Der zentrale Punkt der Methode von Mayring (2015, S. 51) ist das vorher definierte Kategoriensystem, welches in der folgenden Tabelle dargestellt wird.

Tabelle 4.8: Kategoriensystem mit Ober- und Unterkategorien

Oberkategorien	
Oberkategorien Intervention:	***Oberkategorie Outcome:***
► Ursache von pre-existing Bias, die in Daten enthalten sind.	► Behebung von pre-existing Bias, die in Daten enthalten sind.
Forschungsfrage	**Forschungsfrage**
► Was sind die Ursachen von pre-existing Bias in Daten?	► Was ist in der wissenschaftlichen Literatur zur Behebung von pre-existing Bias in Daten bekannt? ► Welche Konzepte oder Modelle werden in den Publikationen thematisiert, die pre-existing Bias in Daten beheben?
Unterkategorien für jeweils Intervention und Outcome	
Datenqualität *Datensammlung* *Datenerhebung* *Datengrundlage*	

Aus Intervention und Outcome, die das Grundgerüst bilden, werden die Oberkategorien Ursache und Behebung gebildet. Die dazugehörigen Unterkategorien sind Datenqualität, Datensammlung, Datenerhebung und Datengrundlage.

Im nächsten Schritt wird die Dokumentenanalyse mithilfe der Ober- und Unterkategorien vorgenommen. Mit der Dokumentenanalyse wird entsprechend den kategorialen Bedingungen Fundstellen identifiziert und kodiert. Das Ziel der deduktiven Analyse ist, mit den vorher festgelegten „Ordnungskriterien einen Querschnitt durch das Material zu legen“ und das Datenmaterial zu strukturieren (ebd., S. 67). In der vorliegenden Arbeit wurde eine „Mischform der qualitativen Inhaltsanalyse“ (ebd.). durchgeführt. Hinzu kam die induktive Kategorienbildung, die sich aus dem Datenmaterial ergeben hat und nicht theoriegeleitet ist. Mayering (2015, ebd.) nennt diese Art der Vorgehensweise „zusammenfassende qualitative Inhaltsanalyse“. Das Ziel der Zusammenfassung ist das Textmaterial so weit zu reduzieren, ohne dabei „den Sinn des Textes zu verfälschen“, um einen Überblick über das zu untersuchende Datenmaterial zu erhalten (ebd.). Die induktiven Fundstellen im Datenmaterial wurden mit einem Code versehen. Im weiteren Verlauf der Analyse wurden ähnliche Aussagen in anderen Dokumenten zu der neu gebildeten Kategorie zugeordnet (ebd., S. 65). Im zweiten Durchlauf wurde das Datenmaterial mit dem erstellten Kategoriensystem überprüft. Die Analyse hat insgesamt 321 Fundstellen in 28 Dokumenten ergeben.[16] Nachfolgend werden die Ergebnisse der Untersuchung dargestellt und interpretierend diskutiert.

[16] Die kodierten Segmente befinden sich als Datenmaterial in einer separaten Word Datei mit 59 Seiten, die auf der beigefügten CD zu dieser Thesis enthalten sind.

5. Darstellung, Diskussion und Interpretation der Forschungsergebnisse

In diesem Kapitel werden die Datenauswertungen im Hinblick auf den theoretischen Bezugsrahmen zusammengeführt und präsentiert. Zunächst werden die Oberkategorien Intervention und Outcome dargestellt. Beide Oberkategorien haben die gleichen Unterkategorien und thematisieren die Qualität, Sammlung, Erhebung und die Grundlage von Daten. Die Unterkategorien, die sich aus dem Textmaterial ergeben haben, werden thematisch strukturiert und erhalten einen Code.

5.1 Oberkategorie – Intervention

In der Oberkategorie Intervention liegt das Erkenntnisinteresse an den Ursachen von pre-existing Bias in Daten und wird von der folgenden Fragestellung geleitet: Was sind die Ursachen von pre-existing Bias in Daten?

Eine der Herausforderungen und Ursachen für die Fehlfunktion von KI-Anwendung ist, wenn eine Diskrepanz zwischen den Daten und der Aufgabenstellung der KI-Anwendung besteht, was zu einer Fehlanpassung führt, weil das Wissen der Welt, so wie sie ist, in den Daten nicht abgebildet ist und Historical Bias genannt wird (Lee et al. 2021, S. 705). Wie im theoretischen Teil aufgezeigt (► vgl. Kap. 3.1.2) erlernen KI-Anwendungen im Lernverfahren aus Vorgängen, die in historischen Datensätzen enthalten sind. Hierbei geben Informationen in den Datensätzen Auskunft zum historischen Verlauf und Resultat einer Situation. Infolgedessen eignen sich KI-Anwendungen das Wissen der Realität an, das in den erfassten historischen Daten enthalten ist. Was zur Folge hat, dass Ein- oder Ausschlussmechanismen einer Gesellschaft (► vgl. Kap. 3.1.2) in kodierte Form zur Geltung kommen, die dazu führen, dass Informationen aus der Vergangenheit, die Gegenwart bzw. Zukunft beeinflussen und zu ungerechtfertigten Gleich- oder Ungleichbehandlung (► vgl. Kap. 3.3) führen. Deutlich wird, dass mathematische Konstrukte an ihre Grenzen kommen, wenn in Daten strukturelle und institutionelle Diskriminierung (► vgl. Kap. 3.1.2) einer Gesellschaft im algorithmischen Entscheidungsprozess einer KI-Anwendung verarbeitet werden und zum Vorschein kommen. In Anlehnung an die Aussage von Zweig (2018, S.15) im Kapitel 2.3 und Umformulierung dieser Aussage wird deutlich, dass Daten von Personen eine bedeutende Rolle für die algorithmische Entscheidung spielen. Denn Algorithmen diskriminieren, wenn sie als Input Trainingsdatensätze bekommen, aus denen sie die Diskriminierungspraktiken einer Gesellschaft erlernen und auf neue und unbekannte Datensätze ihr Wissen anwenden. Der Entscheidungsprozess einer KI-Anwendung führt unter diesem Umstand zur Bevorzugung oder Benachteiligung von sozialen Gruppen,

die die Ressourcenverteilung wie Arbeit, Bildung, Kreditvergabe usw. (► vgl. Kap. 3.1.2) betreffen und die Teilhabechancen erhöhen oder mindern.

Daher besteht die Gefahr, dass KI-Anwendungen die Voreingenommenheit reproduzieren, wenn ein algorithmisches System mit Daten trainiert wird (Zuiderveen Borgesius 2020, S. 4), die historisch bedingte strukturelle Diskriminierung beinhalten. Laut Jo et al. (2020, S. 4) stehen Verzerrungen in historischen Daten für strukturelle Diskriminierung in der Gesellschaft, die sich in den Daten widerspiegeln wie z. B. der historische Mangel an weiblichen Präsident*innen in vielen Ländern oder die Unterrepräsentation von rassischen Minderheiten in Führungspositionen in der Wirtschaft. Demnach würden große wahllose ML-Datensätze in ihrer Gesamtheit Ableitungen und Ergebnisse erzeugen, die diese Verzerrungen widerspiegeln (ebd.). Insbesondere sei die starke Abhängigkeit von historischen Daten laut Cruz Cortés et al. (2020, S. 235) eine wesentliche Ursache für die Reproduktion und Verstärkung unerwünschter Muster. Die zum Training verwendeten Daten seien bereits durch historische Diskriminierung und andere strukturelle Mängel verzerrt, was zu verzerrten ML-Algorithmen führt (ebd.). So macht, laut Leavy et al. (2020, S. 1) das Ausmaß und die Komplexität schädlicher Konstruktionen von *Race* und Geschlecht und deren Einbettung in aktuelle und historische Daten deutlich, wie unvermeidlich es ist, dass Daten eine Form von Voreingenommenheit widerspiegeln oder die Auswirkungen sozialer Ungerechtigkeit erfassen.

Historische Daten, die aufgrund von Repräsentationsverzerrung entstehen, ergeben sich laut Jo et al. (2020, S. 4) aus der Abweichung zwischen der realen Verteilung in einem sozialen Gefüge und dem digitalisierten Datenbestand. Dies könne aus einem ungleichen Zugang zu digitalen Werkzeugen oder soziokulturellen Zwängen resultieren, die eine Digitalisierung oder Datenerfassung verhindern (ebd.). Jüngste Trends hätten laut Leavy et al. (2021, S. 700) aufgezeigt, dass Frauen und PoC Social-Plattformen verlassen, pseudonyme Online-Identitäten annehmen und den Stil sowie Inhalte ihrer Beiträge aufgrund von Belästigungen ändern. Trotzdem würden weiterhin KI-Anwendungen auf der Grundlage von Social-Media-Daten erstellt, ohne dass die Identität derjenigen, die in die Gruppe eingeschlossen oder aus ihr ausgeschlossen werden, umfassend geprüft werden. Dies hat zur Folge, dass geschützte Gruppen ungleich vertreten sind (Kuhlman et al. 2020, S. 2), die zu Representation Bias führen (Lee et al. 2021, S. 705). Repräsentationsverzerrungen entstehen, wenn eine soziale Gruppe über- oder unter- oder überrepräsentiert ist. Infolgedessen führen algorithmische Kodierungen zu einer gruppenübergreifend verzerrten Leistung (Kuhlman et al. 2020, S. 2) . Soziale Zwänge entstehen, wenn der Zugang zu gesellschaftlichen Ressourcen über die soziale Gruppenzugehörigkeit geregelt wird. Wie im Kap. 3.3.1 aufgezeigt, entstehen

bspw. soziale Zwänge bei den Einstellungspraktiken der Mehrheitsgesellschaft, wenn Bewerbende trotz Qualifikation Benachteiligung aufgrund ihrer Herkunftsländer, Religionszugehörigkeit und ihrem Phänotyp erfahren oder automatische Gesichtsanalysesoftware bei Menschen mit dunklem Phänotyp schlecht funktionieren, weil KI-Anwendungen mehrheitlich mit Bildern von *weißen*, männlichen KI-Entwickelnden trainiert wurden (► vgl. Kap. 3.3.2). Diese sozialen Zwänge fließen in die Daten ein und sind die genetischen Informationen einer Gesellschaft, deren Diskriminierungspraktiken durch die algorithmische Informationsverarbeitung einer KI-Anwendung zum Vorschein kommen und die DNA einer Gesellschaft offenbaren.

Ferner können Verzerrungen im Designprozess (► vgl. Kap. 2.3) von KI-Anwendungen entstehen, wenn eine unangemessene Kombination von heterogenen und homogenen Daten in einem einzigen KI-Modell verglichen wird. Bezeichnet wird diese Verzerrung als Aggregation Bias (ebd.). Verzerrungen, die aufgrund von inkorrekten Datenkennzeichnungen oder Kategorisierungen entstehen, werden Measurement Bias (ebd.) genannt. Messverzerrungen können bspw. zu fehlerhaften Rückschlüssen führen, wenn geschützte Merkmale wie Geschlecht, Alter, Religion, Behinderung usw. falsch beschriftet werden oder als Kategorien ungünstig korreliert werden. Daneben können mögliche Leistungsbeeinträchtigungen durch ‚Evaluation Bias' entstehen, wenn Im Designprozess einer KI-Anwendung unangemessene Leistungsmetriken bzw. Richtwerte von anderen KI-Modellen verwendet werden, die nicht die gesamte Population repräsentieren (ebd). Wenn KI-Anwendungen im falschen Kontext implementiert werden, die für die Aufgabenstellung in einer realen Umgebung ungeeignet sind, bspw. als soziotechnische Systeme in die Gesellschaft eingebettet werden, um Ressourcen zu verteilen wie Arbeit oder Sozialleistungen, dann ist die Rede von Implementierungsverzerrung (Deployment Bias) (ebd.).

Solche Implementierungsverzerrungen von KI-Anwendungen sind im Zeitalter von Cloud-Servern und Cloud-Diensten, die sich in einer Datensphäre (► vgl. Kap. 3.1.1 und Kap. 3.1.2) befinden, schwer nachvollziehbar. Die Technologien sind um viele Größenordnungen komplexer als vor den Cloud-Lösungen, deren Komplexität die Interpretierbarkeit schwierig oder unmöglich macht, weshalb Verzerrungen und diskriminierende Entscheidungen festgestellt werden, sobald sie in großem Umfang eingesetzt werden (Acuna et al. 2021, S. 308). Trotz dieser Komplexität und möglichen Gefahr werden laut Bondi et al. (2021, S. 427) utilitaristische Argumente vorgebracht, die die Mängel an Transparenz verschleiern. Die Formulierungen für utilitaristische Argumente hören sich in etwa folgendermaßen an: „optimale Ressourcenallokation und Wohlfahrtsmaximierung innerhalb der Volkswirtschaft" (Datenethikkommission 2019, S. 44) (► vgl. Kap. 1.1). Auch die Zielausrichtung einer KI-Anwendung sei auf den

größtmöglichen Nutzen der Mehrheitsgruppe ausgerichtet, was zur Folge hätte, dass die Auswirkungen auf Minderheiten und Randgruppen leicht übersehen werden können und mit schwerwiegenden Kompromissen verbunden wären, die eine moralische Entscheidung erfordern (Bondi et al. 2021, S. 427). Für moralische Entscheidungen zeigen mathematische Konstrukte in Form von Algorithmen ihre Grenzen auf, weil die Zielausrichtung einer KI-Anwendung auf dem utilitaristischen Prinzip „größtes Glück der größten Zahl“ beruht anstatt „Gerechtigkeit durch Fairness“ (Höffe 2006, S. 5). Somit ist eine utilitaristisch veranlagte KI-Anwendung auf die Maximierung des Allgemeinwohls einer Mehrheitsgruppe in einer Gesellschaft ausgerichtet. Dies liegt in der Natur der statischen Modelle, weil statistische Verfahren die Daten über Populationen klassifizieren. Populationen, die in Datensätzen gut vertreten sind, werden von der „optimale[n] Ressourcenallokation und Wohlfahrtsmaximierung innerhalb der Volkswirtschaft“ (Datenethikkommission 2019, S. 44) profitieren. Populationen, die in Datensätzen falsch oder unvollständig dargestellt werden oder sie mit ungleich stark vertretenen Populationen verglichen werden, hat dieser Umstand weitreichende Auswirkungen auf die Lebensbedingungen und Lebenschancen von sozialen Gruppen oder deren Angehörigen (Scherr 2010, S. 35) (► vgl. Kap. 3.3).

Die schwerwiegenden Kompromissen für Minderheiten und Randgruppen, wie weiter oben beschrieben, gehen zum einen mit struktureller Diskriminierung einher, die die Verteilung von Ressourcen und Gütern betreffen. Feministische Theorien (► vgl. Kap. 3.2) setzen sich seit ihrer Entstehung mit Unterdrückungsmechanismen der strukturellen Diskriminierung auseinander. Ihre Kritik richtet sich an die patriarchalen Machtstrukturen sowie deren Verteilungspraxen der Ressourcen und Güter. Diese Kritik wird u.a. auf den zugewiesenen asymmetrischen Status der Mitglieder einer Gesellschaft ausgeübt, der die strukturelle Diskriminierung hervorruft (vgl. Hooks 2000). Im Zusammenhang mit KI-Anwendungen ist strukturelle Diskriminierung ein Zustand, in dem einer Kategorie von Menschen ein ungleicher Status im Verhältnis zu anderen Kategorien von Menschen zugewiesen wird (Kuhlman et al. 2020, S. 2). Dieser Zustand beschreibt eine mathematische Funktion, die eine Beziehung zwischen zwei Mengen ist. Nennen wir die Elemente dieser Mengen Populationen und Ressourcen. Die mathematische Funktion lautet nun: Stelle eine Beziehung zwischen den Mengen her, wobei die Elemente dieser Mengen Populationen und Ressourcen sind. Durch ein Zusammenspiel dieser Elemente und unbekannten oder falsch modifizierten Variablen entstehen ungleiche Beziehungen in Bezug auf Rollen, Funktionen, Entscheidungen, Rechte und Möglichkeiten, die strukturelle Diskriminierungen aufrechterhalten und verstärken (ebd.). Zum anderen besteht laut Benthall et al. (2021, S. 9) das Problem darin, dass Menschen in KI-Anwendungen als Instrument zur Profitmaximierung

eingesetzt werden. Eine ethische KI-Anwendung wäre dagegen ein soziales System, das die Individuen, mit denen es strukturell verbunden sei, als Zweck und nicht als Mittel behandelt und deshalb würde ein solches System sich selbst und seine einzelnen Mitglieder als freie Lebensform unterstützen:

> „… sustain itself, and its individual members as parts of itself, as a free way of life" (ebd.).

Solch ein gemeinwohlorientiertes System ist noch in weiter Ferne, da die gutgemeinten Absichtserklärungen der großen Big-Tech Konzerne eher ein Lippenbekenntnis sind, weshalb sie an der Unfähigkeit scheitern, sich das Fachwissen und den lokalen Kontext anzueignen, die erforderlich sind, um komplexe soziale Probleme anzugehen (Posada et al. 2021, S. 869). Ihr Interesse gilt den Daten, denn damit haben sie die Macht an sich gerissen, indem sie mit Daten die Machtstrukturen kontrollieren und als Währung einsetzen sowie den Technologiefortschritt maßgeblich vorantreiben (ebd). Ihre neue Rolle bestimmt das Leben von vielen Menschen:

> „… new roles in defining and directing the lives of billions of people" (ebd).

Deshalb wird der Ruf lauter, dass bei der Entwicklung datengesteuerter Lösungen, die Menschen hinter den Zahlen nicht vergessen werden sollten (Bondi et al. 2021, S. 426). Ähnlich wie die Einführung des Frauenwahlrechts (► vgl. Kap. 3.2.1) würde sich das Lippenbekenntnis der Big-Tech Konzerne nur dann ändern, wenn es um Vorteile der eigenen (männlichen) Machterhaltung geht oder eben diese gefährdet sind.

5.1.1 Unterkategorie – Datenqualität

Für das BSI ist die Datenqualität nur dann gegeben, wenn die Trainings- und Testdaten, von denen die KI-Anwendungen ihr Wissen aneignen, aus vertrauenswürdigen Quellen stammen, Qualitätskriterien erfüllen, korrekt kommentiert und angemessen geschützt sind (BSI 2021a, S. 9). Die Fallbeispiele in Kap. 3.3.3 zeigen zum einen auf, dass Rohdaten aus der Datensphäre (► vgl. 3.1.1) wie bspw. soziale Medien ungefragt von Drittanbietenden heruntergeladen werden. Zum anderen zeigt sich, dass Daten von staatlichen Institutionen verwendet werden, die die soziotechnischen KI-Anwendungen einsetzen, um Menschen zu kontrollieren oder sie zu unterdrücken. Insofern bestimmt die gesellschaftliche Wissensinfrastruktur die Datenqualität, in der sich die Daten befinden und entstehen (Leavy et al. 2021, S. 700). Denn sie umfassen eine Ökologie von Menschen, Praktiken, Technologien, Institutionen, materielle Objekte und Beziehungen (ebd.). Ihre Zusammensetzung in Daten beeinflussen die Ergebnisse von Systemen für ML (Jo et al. 2020, S. 1). Eine willkürliche Kategorisierung von Menschen deren Informationen in den Daten enthalten sind, die zum Trainieren von ML-Modellen verwendet werden, kann gefährdeten Gruppen

schaden und gesellschaftliche Vorurteile verbreiten (ebd.). Auf der Plattform GitHub wird seit 2018 eine Liste namens ‚Awful AI' geführt, die alle unfairen KI-Anwendungen erfasst, um Diskussion über die Entwicklung möglicher präventiver Technologien anzuregen bzw. sich zu wehren (David Dao et al. 2022, o.J.). Die Liste ist lang und die auffälligen und unfairen KI-Anwendungen machen sich durch Diskriminierung von Minderheiten und Unterdrückungsmechanismen bemerkbar. Aus der Sicht des Feminismus und der Rassismus kritischen Theorie (engl. Critical *Race* Theories) sind Daten nicht objektiv, neutral oder wertfrei (Leavy et al. 2021, S. 700). Die Daten erfassen Aspekte der Realität, die aus der Sicht derjenigen wahrgenommen werden, die an ihrer Entstehung am stärksten beteiligt sind (ebd.). Denn hinter der Datenerfassung verbergen sich Konzepte und Ideologien, die den Daten für KI-Anwendungen zugrunde liegen (ebd.). Bei der willkürlichen Kategorisierung wird häufig vergessen, dass in den Daten echte Lebensrealitäten von Menschen stecken, die digital in Anerkennungsprozessen (► vgl. 3.1.2) gestaltet und als Zahlen dargestellt werden (Jo et al. 2020, S. 1). Nicht nur methodisches Vorgehen bestimmt die Datenqualität, sondern auch die Hinterfragung der dahinterstehenden Weltanschauung.

5.1.1.1 Code: Diversität

Eine kleine homogene Gruppe von Menschen sind an der KI-Entwicklung beteiligt (► vgl. Kap. 3.3.2), deren Perspektiven und Weltanschauungen sich in die KI-Anwendungen manifestieren. Kuhlman et al. (2020, S. 4 f.) warnen, die verschiedene Datensätze und ihre Verwendungen in KI-Anwendungen untersucht haben, dass mangelnde Perspektivenvielfalt ein kritischer Faktor bei der Entwicklung von KI-Modellen und Methoden gewesen sein könnte, die eine unfaire Voreingenommenheit gegenüber geschützten Gruppen zur Folge hatten. Sie haben festgestellt, dass in allen KI-Bereichen *weiße* männliche Personen mehrheitlich vertreten sind. Durch die einseitige Belegschaft seien fatale Auswirkungen auf die Entwicklung des Forschungsfeldes zu erwarten, wenn es nicht gelänge, die Diversität auf dem Gebiet der Informatik zu verbessern und gleichzeitig Technologien zur Abschwächung von Vorurteilen voranzutreiben, sei ein Scheitern vorprogrammiert, da Forschende und Entwickelnde nicht genügend Ressourcen hätten, um die unfairen Vorurteile, in den von ihnen entwickelten und gebauten Technologien, zu verhindern.

Auch Acuna et al. (2021, S. 311) warnen, dass an wichtigen Schnittstellen wie z.B. KI-relevanten Konferenzen, die ethische Themen behandeln, historisch bedingte und mangelnde Diversität in Bezug auf Geschlecht und *Race* attestiert wird. Wird die erste KI-Konferenz in Dartmouth betrachtet, dann waren alle Beteiligten *weiß* und männlich (► vgl. Kap. 2.1), obwohl Frauen, insbesondere Schwarze Frauen, zu der damaligen Zeit bereits als Mathematiker*innen tätig waren und Daten für die NASA-

Forschung analysierten, jedoch aufgrund von Geschlecht und *Race* diskriminiert wurden (D'Ignazio und Klein 2020, S. 3). Gleiches gilt auch in der aktuellen Zeit. Werden die Mitglieder der Datenethikkommission angeschaut, dann sind sie mehrheitlich *weiß* und männlich, wobei nur sehr wenige Frauen dabei sind und gar keine PoC (DFKI o.J.).

Ebenfalls thematisieren Croeser et al. (2019, S. 425) mangelnde Diversität und ziehen Rückschlüsse auf die vielen Probleme mit bestehenden KI-Bemühungen durch die relative Homogenität des Bereichs. Die fehlenden Perspektiven würden den Fortschritt in der technischen Forschung und bei der Einführung verlangsamen, da im bestehenden Paradigma der KI-Entwicklung das Fachwissen weitestgehend durch eine Reihe von technischen und sozialen Filtern festgelegt werden, die eine breitere Beteiligung verhindern. Diese Faktoren würden es KI-Entwickelnden zwangsläufig erschweren, alternative Sichtweisen auf ihre Arbeit einzuholen, auch von Menschen, die von ihren Projekten negativ betroffen sein könnten.

Zumal KI-Entwickelnde in der Regel nicht in den einschlägigen Methoden geschult werden, wie z. B. in der Identifizierung geschützter Merkmale oder sozialer Kontextfaktoren, die zu Diskriminierung führen können (Galdon Clavell et al. 2020, S. 265).

Diskriminierung liegt allerdings nicht nur an der Schulung von KI-Entwickelnden. Den derzeitigen Stand von Diversität in Unternehmen fasst Chi et al. (2021, S. 449) aus verschiedenen Quellen zusammen und stellt fest, dass u.a. das Thema Diversität eine Art „Happy Talk“ sei. Zwar würden kulturelle Unterschiede in Unternehmen anerkannt und akzeptiert werden, ohne jedoch strukturelle Diskriminierung in Frage zu stellen. Solche strukturellen Diskriminierung wären durch historische und politische Dynamiken geprägt, die beim Thema Diversität jedoch heruntergespielt oder verdeckt werden. Anders sieht es aus, so die Publizierenden, wenn Diversität als Verbesserung der Unternehmensleistung angesehen wird. Diese Art der Anerkennung von Diversität sei eine rücksichtslose Sichtweise, in der der Wert von unterrepräsentierten Menschen an ihren Mehrwert für die Mitglieder der Gruppe gebunden sei …

> „… predatory view (…) in which he worth of underrepresented people is tied to their value add to in-group members”. (ebd.).

Dass diese Sichtweise kein amerikanisches, sondern auch ein deutsches Phänomen ist, wird durch die „Charta der Vielfalt für Diversity in der Arbeitswelt“ deutlich. Dort heißt es:

> „Die Diversität der Mitarbeitenden mit ihren unterschiedlichen Fähigkeiten und Talenten eröffnet Chancen für innovative und kreative Lösungen (…) Die Anerkennung und die Förderung vielfältiger Potenziale schaffen wirtschaftliche Vorteile für unsere Organisation.“ (Charta der Vielfalt e.V. o.J.)

Wird der Vorstand angeschaut oder gar die Mitglieder dieser Vereinigung, dann sind es mehrheitlich Menschen aus der Mehrheitsgesellschaft. Diversität wird zwar als zentrale Transformation für die Gesellschaft betrachtet, jedoch keine Mechanismen für Inklusion etabliert. Diversität bleibt ein Lippenbekenntnis und Inklusion ist eine Frage der Teilhabe wie im Folgenden deutlich wird.

5.1.1.2 Code: Inklusion

Inklusion in Datentechnologien bedeutet für Kuhlman et al. (2020, S. 1), dass unterrepräsentierte Gruppen von Personen, die ethnischen Minderheiten oder Geschlechtergruppen angehören und in der Vergangenheit möglicherweise Diskriminierung erfahren haben, in Bezug auf ihre Beteiligung an der technologischen Entwicklung unterrepräsentiert sind. Wohingegen mit Diversität auch gefährdete Bevölkerungsgruppen gemeint sind, die nicht über das notwendige soziale Kapital verfügen, um sich selbst zu vertreten. Darunter lassen sich u.a. Kinder, inhaftierte Personen, Studierende und wirtschaftlich Bedürftige usw. zuordnen. Laut der Publizierenden zeigen die Daten dieser gefährdeten und unterrepräsentierten Gruppen, dass sie nach wie vor systembedingten strukturellen Verzerrungen ausgesetzt seien, die sich schon bei der Datenerfassung zeigen würden.

Wobei Chi et al. (2021, S. 449) argumentieren, dass Inklusion in Datentechnologien ohne entsprechende Machtausstattung die Wahrscheinlichkeit von Datengewalt erhöhen kann, indem Inklusion ohne Macht unterdrückerische strukturelle Bedingungen normalisiert. Mit Machtausstattung ist wahrscheinlich die Teilhabemöglichkeiten an Prozessen gemeint. Den auch Teilhabemöglichkeiten bedeuten zwangsläufig nicht, dass die Beteiligung an Prozessen mit Entscheidungsbefugnissen einhergehen. Teilhabemöglichkeiten haben Stufen der Beteiligungsart, die sich in ihrer *Machtausstattung* unterscheiden. Im Partizipationsmodell von Straßburger et al. (2014, 232 f.) hat Teilhabe verschiedene Stufen. In der Vorstufe bedeutet Partizipation aus der Perspektive der Bürger*innen, dass sie sich informieren können, dann im Vorfeld von Entscheidungen hierzu Stellung nehmen dürfen und im letzten Schritt können sie zwar Beiträge vorbringen, jedoch ohne Garantie, dass diese Beiträge in die Entscheidungen einfließen (ebd.). In den Vorstufen des Partizipationsmodells werden lediglich Meinungen erfragt, die auch versteckte Marketingstrategien sein können, um die Akzeptanz zum Themengegenstand zu erhöhen. Erst in der nächsten Stufe des Partizipationsmodells ist die Beteiligung an Entscheidungsprozessen möglich, indem Bürger*innen mitwirken und gemeinsam mit Fachkräften an Problemlösungen arbeiten (ebd.). In den darauffolgenden Stufen erweitert sich die Teilhabe an Entscheidungsprozessen, die die Selbstverantwortung und Entscheidungsfreiheit beinhalten. Die oberste Stufe des Partizipationsmodells ist die

zivilgesellschaftliche Eigenaktivität, die die Bürger*innen befähigt sich selbst zu organisieren und ihr Vorhaben eigenständig umzusetzen (ebd.). Partizipation aus der institutionell-professionellen Perspektive bedeutet, dass je nach Partizipationsstufe der Grad der Bürger*innen-Befähigung erhöht wird (ebd.). Also ist zu fragen, welche Stufe der Partizipation gemeint ist, wenn Machtausstattung gefordert wird.

5.1.1.3 Code: Ursachen von Diskriminierung

Vorfälle von Diskriminierungen in KI-Anwendungen haben laut Leavy et al. (2021, S. 696) bereits aufgezeigt, dass ein großer Anteil der automatisierten Entscheidungsfindungsprozesse auf menschliche Vorurteile in Daten zurückgeführt werden können. Die Voreingenommenheit als ein nicht zu vermeidendes Merkmal von Daten, die im Rahmen menschlicher Prozesse erhoben werden, wird inzwischen allgemein anerkannt. Wie im Kap. 3.3.1 erörtert, können die zugrundeliegenden Trainingsdaten ihren Ursprung in der Gesellschaft insgesamt, in Subkulturen und in formellen oder informellen, privaten oder öffentlichen Organisationen und Institutionen haben (Friedman et al. 1996, S. 333). Diese Vorurteile kommen bei der Erstellung, Sammlung und Kommentierung von Datensätzen zum Vorschein. Des Weiteren sind die Vorurteile das Ergebnis von historischen Ungerechtigkeiten. Aufgrund dessen würden Verzerrungen bei der Darstellung, Klassifizierung und Kategorisierung durch Menschen in personenbezogenen Daten entstehen. Solche diskriminierenden Praktiken münden in rassifiziertem Profiling, wenn Klassifizierungen aufgrund von angenommenen Eigenschaften gebildet werden. Daher sei die Kategorisierung und die unterschiedliche Behandlung von Menschen, die ein Recht auf Privatsphäre sowie Nichtdiskriminierung haben, grundlegend untergraben.

Um dieses Dilemma zu lösen, hat sich die Informatik laut Posada et al. (2021, S. 866) auf die Frage konzentriert, wie Algorithmen „fair“ und „unvoreingenommen“ gestaltet werden können, und Wege erforscht, um vielfältigere Daten zu sammeln und gleichzeitig die algorithmische Diskriminierung in mathematischen Modellen zu lösen. Für solche mathematischen Modelle ist es unerlässlich, dass die Daten je nach Kategorie oder Untergruppe unterschiedlich behandelt werden müssten. Allerdings würde die rein statistische und mathematische Betrachtung von „Fairness“ den gesellschaftlichen Implikationen der KI-Anwendung nicht gerecht. Diese prädiktiven Systeme seien von Natur aus soziotechnisch, weshalb die Konstruktion und Implementierung von Algorithmen ein Verständnis von sozialen Kontexten und Beziehungen erforderlich macht, das über Rechenmodelle hinausgeht.

So konstatieren Leavy et al. (2020, S. 2), dass mit der zunehmenden Erkenntnis von Voreingenommenheit und Diskriminierung in KI-Anwendungen, diese von Daten ausgehen. Unter diesem Aspekt sollte ihnen nach die Bedeutung der menschlichen

Arbeit an Daten und der Nutzen einer Datenkuration für KI-Anwendungen stärker in den Vordergrund treten. Deshalb fordern die Publizierenden den Mythosabbau in Bezug auf die Objektivität von Algorithmen, die Fairness implizieren.

Um das Repräsentationsproblem in der Belegschaft zu umgehen, könnten die Unternehmen laut Chi et al. (2021, S. 456) Ziele der Diversität und Inklusion in Form von diversen Datensätzen oder personalisierten Tools vorantreiben, die unterschiedliche Verhaltensweisen und Kontexte berücksichtigen. Auch, wenn diese Ziele und Arbeitsabläufe zu besseren Kundenerfahrungen führen können, sei die Änderung von Datenpraktiken möglicherweise einfacher als die Änderung von Einstellungspraktiken, Unternehmenskulturen und Geschäftsmodellen als die Beseitigung von struktureller Diskriminierung.

5.1.1.4 Code: Fairness

In den Anfängen der KI-Entwicklungen, die laut Posada et al. (2021, S. 866) stark von der Philosophie beeinflusst waren, konzentrierten sie sich auf allgemeine ethische Grundsätze, die KI-Systeme befolgen sollten. Diese Grundsätze wie Transparenz und Fairness wurden in erster Linie von dominanten Akteur*innen aus der Industrie und der Politik definiert. Allerdings blieben diese Prinzipien abstrakt und ohne klaren Konsens darüber, wie sie umgesetzt werden sollten, so dass zielführende Diskussionen über Politik und Regulierung auf der Strecke blieben. Darüber hinaus wurden diese Diskussionen wegen ihrer mangelnden Anwendbarkeit und der Finanzierung durch Big-Tech Konzerne als „ethics washing“ kritisiert.

Was zur Folge hat, dass die Umsetzung von Fairness in KI-Anwendungen eine Herausforderung bleibt und wie im Kap. 1.1 erörtert, der Forschungsstand zur Bewältigung bestehender ethischer Herausforderungen von unvoreingenommen und diskriminierungsfreien KI-Anwendungen noch weitestgehend sich im Anfangsstadium befindet (Zhang et al. 2022, 17 ff.). Fairness durch mathematische Konstrukte in Algorithmen nachzubilden, die gesellschaftliche Fairnessmaße berücksichtigen sollen, ist dem Umstand geschuldet, dass unterschiedliche Definitionen von Fairness und Bias existieren und der KI-Welt im Allgemeinen und beim ML im Besonderen als Problem im Modellieren von KI-Anwendungen begegnen (Leavy et al. 2021, S. 696). Durch das heuristische Vorgehen wird die KI-Forschung von der Erkenntnis begleitet, dass es viele verschiedene Bedeutungen von „Fairness“ gäbe, die unterschiedliche Kompromisse implizieren und eigentlich Sensibilität für Kontext und Weltanschauung erfordern (Benthall et al. 2021, S. 7). So wurde bspw. laut Kuhlman et al. (2020, S. 6) beobachtet, dass traditionelle eurozentrische Erkenntnistheorien in Forschungsgemeinschaften oft von den kulturellen Praktiken unterrepräsentierter und gefährdeter Gemeinschaften abgekoppelt sind und in Ignoranz münden.

Genauso wie in Kapitel 3.3.3 aufgezeigt, wird von KI-Anwendung mit ML erwartet, nicht aus Daten die gesellschaftlichen Vorurteile zu erlernen, die mit patriarchalen Machtstrukturen, neoliberalem Kapitalismus oder dem Erbe des Kolonialismus einhergehen, jedoch die Ansprüche der Fairness erfüllen sollen, um gerechte Entscheidungen zu treffen.

Um die Ignoranz mit mathematischen Mitteln zu lösen, sind z.B. zur Risikominderung einige Fairness-Toolkits entstanden, die die vorhandene Unfairness in Daten durch „De-Biasing-Techniken“[17] in den Entwicklungsphasen eines KI-Modells aufheben (Lee et al. 2021, 705 ff.). So seien zwar diese Abhilfemaßnahmen in bestimmten Kontexten nützlich, jedoch sollten laut den Publizierenden Abhilfemaßnahmen bspw. eher bei Menschen und Prozessen als bei KI-Modellen selbst angewendet werden, z. B. bei der Schulung von menschlichen Datenkennzeichnenden oder die Sammlung vielfältigerer Datensätze. Die De-Biasing-Techniken würden davon ausgehen, dass Fairness mathematisch operationalisiert werden könne. Daher würde diese Ansicht oft kritisiert, weil sie den gesellschaftlichen und historischen Kontext außer Acht lassen würden. Darüber hinaus besteht die Schwierigkeit darin, dass es keine systematisierte Methode zur Identifizierung und Abschwächung von unfairen KI-Modellen gibt. Wie im Kapitel eins dargestellt, ist mit dem AIC4 Kriterienkatalog ein Leitfaden zur Risikoanalyse entstanden, die die technischen Gegebenheiten rund um Hardware und Software untersucht. Jedoch werden im AIC4 keine Empfehlung für Bias ausgesprochen.

Ferner hätten KI-Entwickelnde Schwierigkeiten mit Voreingenommenheit und 'blinden Flecken', die bei den Menschen in der ML-Entwicklungspipeline vorhanden sein können, wie z. B. bei Crowdworkern oder Studienteilnehmende, explizit zu berücksichtigen. Diese würden bei den mathematischen Tests der KI-Modelle nicht erkannt werden und würden daher eine qualitative Identifizierung erfordern.

5.1.1.5 Code: Diskriminierung von sozialen Gruppen

Eine neue Erscheinungsform der Diskriminierung ist die indirekte Diskriminierung, die laut Zuiderveen et al. (2020, S. 5), dann in Erscheinung tritt, wenn eine Praxis, die scheinbar neutral erscheint, jedoch im Entscheidungsprozess einer KI-Anwendung aufgrund geschützter Merkmale Menschen diskriminiert. Indirekte Diskriminierung wird in den Vereinigten Staaten als "disparate impact", wohingegen direkte Diskriminierung als "disparate treatment" bezeichnet wird.

[17] Wird im Kapitel Outcome als Lösungsmethode vorgestellt.

Diese indirekten Diskriminierungsformen kritisiert die LGBTQ+-Community und wirft laut Tomasev et al. (2021, S. 260) den KI-Entwickelnden mangelnde Perspektivenvielfalt vor. Sie fordert von KI-Forschenden Lösungsvorschläge ein, die Fairness in Bezug auf die queere Community ausdrücklich berücksichtigen sollen. Während beobachtbare Merkmale (wie z. B. das gesetzliche Geschlecht, das Einkommen oder der Beruf) und unveränderliche Merkmale (wie z. B. die ethnische Gruppe) für die Trainingsdaten in KI-Anwendungen gemessen und aufgezeichnet werden können, sind im Gegensatz dazu Merkmale wie sexuelle Orientierung und Geschlechtsidentität häufig nicht messbar bzw. beobachtbar. Dieser Datenmangel entstünde möglicherweise aufgrund von Datenerhebungen, die die Queer Community nicht berücksichtigt. Die Berücksichtigung der sensiblen Daten kann jedoch u.U., wenn Missbrauchsabsichten vorhanden sind, die Privatsphäre oder die Sicherheit einer Person bedrohen. Der Versuch, solche Informationen zu kategorisieren, zu etikettieren und aufzuzeichnen, könne von Natur aus unzureichend sein und zu direkten oder indirekten Diskriminierung führen, weil einige Merkmale unbeobachtet bleiben, da sie grundsätzlich nicht messbar seien. Diese Unstimmigkeiten bei der Wahrnehmung und Messbarkeit würden zu Diskrepanzen und Spannungen führen. Daher sei es wichtig, bei der Analyse der Fairness von KI-Anwendungen einen angemessenen Granularitätsgrad anzuwenden.

Auch hinsichtlich des Alters weisen Daten Qualitätsmängel auf. So haben Park et al. (2021, S. 835) festgestellt, dass nur 26 Prozent der 92 Gesichtsbilddatensätze altersbezogene Metadaten enthalten. Darüber hinaus seien die Normen für die Angabe des Alters der Probanden in diesen Datensätzen sehr uneinheitlich, wobei einige Datensätze das Alter einfach in einer binären Kategorie von "jung" und "alt" dokumentieren, während andere ungleichmäßig verteilte Kategorien verwenden, bei denen die älteren Alterskategorien einen viel größeren Bereich umfassen als die Jüngeren. Darüber hinaus würden in den Altersmetadaten dieser Datensätze ältere Erwachsene nur selten berücksichtigt, wobei die höchste Altersklasse oft mit 50 Jahren oder sogar darunter endet. Die wenigen Ausnahmen seien spezielle Datensätze, die zum Trainieren von Algorithmen für die Altersschätzung erhoben wurden, aber selbst in diesen Datensätzen enthielt die Altersverteilung nur eine kleine Anzahl älterer Erwachsener und sehr wenige (oder gar keine) noch ältere Erwachsene. Schlussendlich stellen sie noch fest, dass selbst in den Datensätzen, in denen das Alter enthalten war, diese Information oft weder verifiziert noch direkt von den Probanden bezogen waren, sondern von Crowdworkern kommentiert wurden, die das Alter der Probanden anhand ihres Aussehens vermuteten oder es aus öffentlich zugänglichen Informationen ableiteten.

Die Daten von Kindern sind ähnlich mit Qualitätsmängel behaftet. So weisen Bryant et al. (2019, S. 381) darauf hin, dass Technologien zur Gesichts- und Emotionserkennung für Bildungseinrichtungen in Betracht gezogen werden, um sicherheitsbedingte Probleme durch Überwachung zu lösen. Die Ergebnisse ihrer Untersuchungen mit fünf Datensätzen, in denen die Gesichtsemotionen von Kindern enthalten sind und in acht Emotionserkennungssystemen zum Einsatz kamen, zeigen auf, dass der Einsatz von KI-Anwendungen unbestreitbar verfrüht seien. Daher sei dies ein unmittelbares und dringendes Problem. Wenn die Emotionserkennungssysteme nicht ganzheitlich für die Zielgruppen konzipiert werden, würden indirekte Voreingenommenheit und Ungerechtigkeit in diesen Systemen fortbestehen, mit potenziell verheerenden Auswirkungen.

Für den Einsatz von KI-Anwendungen wurden durch den Obersten Gerichtshof in den USA die Möglichkeit eingeräumt, im Auswahlverfahren der Hochschulzulassungen Aufsätze von Bewerbenden unterstützend zu bewerten. Diese Entscheidung hat Alvero et al. (2020, S. 202) veranlasst, die Bewerbungsaufsätze als Textdatensätze zu untersuchen, um herauszubekommen, inwieweit auf demografische Merkmale der Bewerbenden durch ML Rückschlüsse aus den Aufsätzen gezogen werden können. Dabei haben sie festgestellt, dass die statistischen Modelle aufgrund der Daten das Geschlecht und Haushaltseinkommen mit hoher Genauigkeit vorhersagen konnten, was zu Bevorzugung oder Benachteiligung von Bewerbenden führen könnte.

Textanalysen in Daten sei laut Leavy et al. (2021, 696 ff.) eine besondere Herausforderung für KI-Anwendungen, denn die menschliche Sprache sei von stereotypischen gesellschaftlichen Konzepten durchdrungen. Daher fragen sich die Publizierenden, wie rassistische, frauenfeindliche und andere diskriminierende Inhalte bei einer Datenkuration erkannt werden sollen. Denn die zentrale Prämisse rassenkritischer Ansätze behandle den Rassismus als anpassungsfähig, weil es eine Ideologie sei, die sich selbst auflöst, indem er sich von verschiedenen Metaphern und Konzepten bediene und oft durch ständige Debatten funktioniere. Solche Hassreden seien zwar in Europa illegal, aber es bestünde wenig Einigkeit darüber, welche Arten von Inhalten rassistisch und frauenfeindlich sind oder andere Formen des sozialen Hasses enthalten.

5.1.2 Unterkategorie – Datensammlung

Der Prozess der Datensammlung und Datenkuration für Algorithmen des ML ist vom Standpunkt des Feminismus oder der Rassismus kritischen Theorie aus, laut Leavy et al. (2020, S. 2) niemals objektiv. Die algorithmische Datenverarbeitung, die den Gesetzmäßigkeiten von mathematischen Funktionen unterliegen (► vgl. 2.3), sind

insoweit objektiv, wie sie aus den ihnen zur Verfügung gestellten Daten sich das Wissen aneignen. Ihre Objektivität endet dort, wo Datensätze Menschen und deren Verhalten repräsentieren. Wenn Datensammlung nicht die geschützten Merkmale wie ethnische Herkunft, Geschlecht, Religion und Weltanschauung, Behinderung, Alter und sexuelle Identität beinhalten darf, um die Privatsphäre zu schützen, was bleibt dann übrig? Wie sollen Menschen dann voneinander unterschieden werden, um differenzierte Datensätze zu erhalten? Selbst durch Namen können Rückschlüsse auf ethnische Herkunft, Geschlecht, Alter oder Religion gezogen werden oder Adressen von Privatpersonen können Anhaltspunkte auf sozioökonomische Verhältnisse bieten. Daher wird jede Datensammlung in Übereinstimmung mit einer Weltanschauung oder Ideologie durchgeführt, die die geschützten Unterscheidungsmerkmale von Menschen sowie deren Verhalten in Datensätzen abspeichert und als Trainingsdaten für KI-Anwendungen zum Erlenen der Unterscheidungsmerkmale zur Verfügung stellt (ebd.). Die Datensammlung und Datenkuration ist demnach ein politischer Akt, der durch KI-Anwendungen durchgesetzt wird (ebd.).

Es werden laut Schelenz et al. (2021, S. 905) umfangreiche und sensible Daten gesammelt, ohne dass die Bedingungen für eine informierte Zustimmung bekannt sind. Solche Daten werden für personalisierte Werbung verwendet, indem die Schwachstellen der Anwendenden ausgenutzt werden, was einer Manipulation gleichkommen würde und die freie Entscheidungsfindung der Anwendenden behindern kann.

Nach Kuhlman et al. (2020, S. 1) legen die Datensätze von schutzbedürftigen und unterrepräsentierten Gruppen dar, dass sie nach wie vor systembedingten strukturellen Verzerrungen ausgesetzt sind und das Ergebnis automatisierter Entscheidungsfindungsprozesse verfälschen. Diese Verzerrungen betreffen nicht nur diejenigen, die in den vorangegangenen Kapiteln aufgezeigt wurden, sondern alle, die nicht in der KI-Community vertreten oder deren Mitgliedern ähnlich sind. Wie in den Kapiteln zuvor aufgezeigt, sind in der KI-Community mehrheitlich *weiße*, junge, männliche Forschende tätig und als …

> "… a small fraction of humanity is currently engaged in the development of a set of technologies that are already transforming the everyday lives of almost everyone else." (Floridi et al. 2021, S. 12).

Soziale Ungerechtigkeiten, die durch KI-Anwendungen hervorgerufen werden, haben die Sozialwissenschaft auf den Plan gerufen. Sie erforschen mittlerweile die Machtverhältnisse, die hinter der Entwicklung und dem Einsatz von KI-Anwendungen stehen, so Posada et al. (2021, S. 866). Ihnen nach, befasst sich ein Teil der Sozialwissenschaften mit den soziotechnischen Aspekten der Datenlieferketten, die die Entwicklung des ML ermöglichen. Die Untersuchung dieser Datenlieferkette umfasst den

Datenerfassungsprozess für Eingabe-, Trainings- und Feedbackdaten sowie die Fragen rund um den Datenschutz bei der Datensammlung, aber auch alle beteiligten Akteur*innen der Datenversorgung bspw. Datenbearbeitenden, Datenbroker*innen, institutionelle und staatliche Akteur*innen, usw.

5.1.2.1 Code: Ökonomie

Posada et al. (2021, 865 ff.) konstatieren, dass die größten Akteur*innen im KI-Bereich gerade einmal neun Unternehmen mit Sitz in China und den USA seien. Ihnen nach hätte der unersättliche Datenhunger von sozialen Medien, Suchdiensten und Plattformanbietenden eine zwanghafte Art die Nutzenden an sich zu binden, denn ohne diese würden die Unternehmen keine vorteilhaften Positionen beim Datenzugang haben. So würden sich solche Unternehmen durch die Ökonomie der Datafizierung auf die finanziellen Imperative und Gewinne konzentrieren, die Daten aus der beobachtbaren Realität erzeugen. Die sozialen Auswirkungen der Datensammlung würden die KI-Entwickelnden und Technologieunternehmen eher mit Passivität begegnen, denn nach ihren Vorstellungen sind die Daten ein nicht rivalisierendes Gut, das nicht durch Nutzung verbraucht wird und ein Nebenprodukt der ökonomischen Aktivität ist. So sei die Auffassung verbreitet, dass mit Daten als Wirtschaftsgut gehandelt werden dürfe.

Die finanziellen Imperative und Gewinne rufen die Regierungen auf den Plan, um durch die Konzeptualisierung von Daten ökonomisches Wachstum zu generieren, wenn bspw. Daten als „die Rohstoffe des 21. Jahrhunderts“ bezeichnet werden (Die Bundesregierung 2016, o.S.). Mit der Aussage steht jedoch der Wachstumsanspruch der westlichen Regierungen im Widerspruch zum Grundrecht des Einzelnen auf Privatsphäre und führt zu einem Paradoxon der Ethik- und Fairnessansprüche, weshalb Kak (2020, S. 309) die Diskussion um die universalisierten Paradigmen der Ethik und Fairness des Westens kritisiert. Ihr nach wiederholen sich die kolonialen Ungerechtigkeiten und die anhaltende Unterordnung der Länder des Globalen Südens, wenn die globale Versorgungskette der KI-Anwendungen betrachtet wird. Demnach umfasst die Versorgungskette u.a. personenbezogene Daten, natürliche Ressourcen wie Edel- und Schwermetall für die Herstellung von Hardware, billige Arbeitskräfte usw., womit bspw. laut der Autorin auch die indische Regierung die Konzeptualisierung von Daten als nationales Eigentum voranbringen will, um den ehrgeizigen Wachstumsanspruch der indischen Industrie voranzubringen.

5.1.2.2 Code: Überwachung

Mit der Datafizierung, die als Wirtschaftsgut angesehen wird, eröffnen sich Wege zur Überwachung und Informationskontrolle. Datafizierung sei Posada et al. (2021, 866 ff.) zum einen aus der politischen und ethischen Perspektive ein Phänomen der

„Datenüberwachung“, welche auch als die *Disziplinierungs- und Kontrollpraxis zur Überwachung, Aggregation und Sortierung* bezeichnet wird. Im Gegensatz zur physischen Überwachung, beobachtet die Datenüberwachung den Schatten, den die Person bei der Durchführung von Transaktionen wirtschaftlicher, sozialer oder politischer Natur wirft. Die Technologien, die die Datenüberwachung ermöglichen, würden nach den Publizierenden immer stärker in das Leben des Einzelnen eindringen, wobei dies mit Sicherheitsbedürfnissen begründet sei. Zum anderen ist das Aufkommen der Datafizierung mit dem Buchdruck vergleichbar. Diese Innovationen hätten Akteur*innen hervorgebracht, die aus der grundlegenden Veränderung des Umfelds Kapital geschlagen hätten, wie z.B. das Unternehmen Google, zum allgegenwärtigen Hüter von Informationen wird und die Suche kontrollieren würde. Zusammen mit Social-Media-Giganten wie Facebook, Tiktok und Twitter und den App-Gatekeepern wie Apple und Amazon würden sie effektiv das Wissen kontrollieren. Auch zu den Zeiten des Buchdrucks wurde begrenzt und bestimmt, welche Informationen in welchem Format und von wem empfangen werden durften. So hätten die europäischen Kolonialherren in Lateinamerika das gedruckte Material reguliert, um die Verbreitung von Informationen und Ideen zu kontrollieren. Was zu der Schlussfolgerung führt, dass durch die Datafizierung die Möglichkeit geschaffen wird, den Zugang zu Ressourcen durch Unterdrückungsmechanismen (Hill Collins 2000) in Form von Überwachung und Informationskontrolle zu regeln, die bereits in Kap. 3.3.2 durch die Beispiele aufgezeigt wurden.

Unterdrückungsmechanismen werden mit Argumentationen zur Fairnesseinhaltung vorgebracht, die laut Cooper et al. (2021, S. 51) den Vorschlag unterbereiten, noch mehr Daten über die unprivilegierten Gruppen zu sammeln, um die Klassifikationsunterschiede zwischen den Gruppen anzupassen und, dabei die Daten der privilegierten Gruppe unangetastet zu lassen. Was also nicht passt, soll nach diesem Vorschlag passend gemacht werden oder mit den Worten von Scherr (2010, S. 43) werden Unterscheidungspraxen angeglichen, indem die, ‚die in unerwünschter Weise anders‘ sind, noch mehr Daten liefern sollen, um die Klassifizierungsergebnisse von KI-Anwendungen zu verbessern, wobei sich die ‚die Normalen‘ der Überwachung entziehen dürfen. Die Daten der privilegierten Gruppen sind die Norm, während die Daten der unprivilegierten Gruppen eine Abweichung von dieser Norm darstellen. Die eigene privilegierte Position in der Gesellschaft, die mit struktureller Bevorzugung einhergehen, wird bei diesem Vorschlag gar nicht hinterfragt.

Dadurch würde sich laut Cooper et al. (ebd.). die Last der gerechten Klassifizierungsergebnisse einer KI-Anwendung, ungerechterweise auf die unprivilegierte Gruppe verlagern, während die bereits privilegierte Gruppe zusätzlich

privilegiert wird. Diese Annahme sei laut den Publizierenden eindeutig falsch. Es sei weithin anerkannt, insbesondere in der soziotechnischen Literatur, dass die Datenerhebung und -sammlung häufig eine Form der Überwachung ist. Sie sei kein neutraler Akt und würde in den USA im Allgemeinen nicht gleichmäßig auf alle demografischen Gruppen angewandt. Dies hätte sich schon im Hinblick auf die historischen Überwachungspraktiken von nicht-*weißen* und queeren Personen in den USA gezeigt.

Ähnliche Vorschläge werden auch bei Gesichtserkennungssystemen gemacht. Damit die Genauigkeit von Gesichtserkennungssystemen verbessert werden kann, wird laut Leavy et al. (2021, S. 697) vorgeschlagen, den Anteil von Menschen mit dunklem Hauttyp in den Datensätzen zu erhöhen. Was zunächst fair klingt, könne im breiteren Kontext zur Aufrechterhaltung sozialer Ungerechtigkeit und zur Verschärfung bestehender sozialer Diskriminierung führen. Die Forderung nach erhöhter Datenerfassung und -sammlung über benachteiligte Gruppen könne die soziale und wirtschaftliche Ungleichheit aufrechterhalten.

Wobei Whittlestone et al. (2019, S. 199) der Meinung ist, dass Unternehmen und Regierungen personenbezogene Daten nutzen können, um die Nachrichten, Angebote und Dienste, die den Menschen angezeigt werden, zu personalisieren. Diese Personalisierung könne es den Menschen erleichtern für sich die richtigen Produkte und Dienstleistungen zu finden, aber eine derart feingliedrige Unterscheidung zwischen den Menschen könne die gesellschaftlichen Ideale der Staatsbürgerschaft und Solidarität gefährden.

5.1.3 Unterkategorie – Datenerhebung

Mit dem Beginn des Internetzeitalters sind durch die Datenerhebung Möglichkeiten entstanden, gesellschaftliche Vorgänge in digitales Wissen umzuwandeln und zu verwalten. Die Datensphäre hat sich zu einem Universum der Datafizierung verwandelt und die dazugehörigen Technologienentwicklungen sind kleiner, günstiger und effizienter geworden (► vgl. Kap 3.1.2), womit mehr Möglichkeiten für Gesellschaften entstanden sind (Posada et al. 2021, S. 865). Mit dem Aufkommen des ML sind jedoch mit der Datenerhebung auch die Möglichkeiten entstanden, zu bestimmen, welche Informationen von wem eingesehen werden (ebd.). Algorithmen des ML haben sich zu Schiedsrichter*innen von Wahrheit und Macht entwickelt, die Menschen in verschiedene Filterblasen schieben, was zur Folge hat, dass innerhalb einer Gesellschaft eine tiefe Kluft entstehen kann (ebd.).

5.1.3.1 Code: Liberalismus

Bei der Datenerhebung werden in westlichen Demokratien liberale Gesetze angewandt, die laut Benthall et al. (2021, 3 f.) scheitern, weil die Gesetzesanwendung mit ethischen Grundsätzen für KI-Anwendungen sowie den Datenschutz und Privatsphäre inkonsistent seien. Denn nach den Publizierenden stehen die Wahrung der Autonomie des Einzelnen im Widerspruch zu wachstumsorientierten Märkten, die aus personenbezogenen Daten Profit erzeugen wollen sowie KI-Anwendungen, die gleiche Rechte für alle gewährleisten sollen, in dem sie Ressourcen und Güter fair verteilen. Gleiche Rechte für alle entspringt dem egalitären Impuls des neuen Liberalismus, der laut Posada et al. (2021, S. 867) in der KI-Ethik unter dem zentralen und umstrittenen Begriff „Fairness“ sehr lebendig sei. Fairnessansprüche, die den Konzepten der Gerechtigkeitstheorie (► vgl. Kap. 3.3.3) entsprechen sollen, bringen die Weltanschauung des Liberalismus ins Wanken, denn laut Benthall et al. (2021, 3 f.) sind die Regulierungsmaßnahmen und ethischen Anforderungen mit den politischen und ökonomischen Dimensionen weitestgehend unvereinbar. In ihrer Analyse kommen die Publizierenden zu der Erkenntnis, dass eine KI-Anwendung insoweit ethisch sein wird wie der Zweck des sozialen Systems, das sie einsetzt und betreibt (ebd.). Demzufolge steht die Profitmaximierung der freien Märkte, die die Nutzenden als Mittel ansehen, im Widerspruch zu sozialen Gerechtigkeit. Die Theorien des Liberalismus und der sozialen Gerechtigkeit besagen, dass die Voraussetzung für die Legitimität eines Systems die Zustimmung derjenigen erfordere, die dem System unterworfen sind (Acuna et al. 2021, S. 308). Übertragen auf die KI-Ethik würde die Argumentation lauten, dass KI-Anwendungen die Bestätigung derjenigen suchen sollten, die am meisten von ihnen betroffen sind (ebd.). Wie bereits in Kapitel 2.1 aufgezeigt, sind in den medialen Diskursen die Interessen der Wirtschaft dominant vertreten, während die Interessen der Zivilgesellschaft oder politisch Agierenden kaum vertreten sind (Fischer et al., S. 6, 2021). Daraus resultierend wird deutlich, dass die Bestätigung derer, um die es geht, nicht berücksichtigt werden und wirtschaftliche Interessen eine hochrangige Priorisierung erfährt, wobei die Legitimität stillschweigend angenommen wird.

Was jedoch verlangt wird, ist laut Posada et al. (2021, 866 ff.), dass Anwendende die Art und Weise, wie sie ihr Lebensstil gestalten und das Verständnis ihrer Rechte sowie den Umgang mit ihren Daten neu zu bewerten. Das gesamte System der Datafizierung von Kommunikation, über soziale Interaktionen bis zum Wissenstransfer findet auf einem völlig neuen Medium statt, das nicht dasselbe ist wie die Medien, mit denen Menschen vertraut sind. Der Übergang zur Datafizierung und der Einsatz von KI, würde sich besser verstehen lassen, wenn der Übergang von der Handschrift zum Buchdruck betrachtet wird. So seien vor dem Buchdruck handgeschriebene Texte

buchstäblich heilig gewesen, weil nur wenige lesen konnten. Heute gilt dasselbe für Daten. Daten seien geheimnisvoll und eine Domäne von Fachleuten, denn sie wären es, die die heute weit verbreiteten Instrumente entwickelten. Der Haken an der Sache sei dabei, dass Einzelne das Durchdringen von Datafizierung in ihr Lebensumfeld zulassen und bei diesem technologischen Wandel einen Teil von sich selbst preisgeben.

5.1.3.2 Code: Datenmonetarisierung

Wenn schon der Alltag der Menschen von Technologieunternehmen kolonisiert wird, werden Überlegungen angestellt, die entstandenen Daten zu monetarisieren. Dazu gibt Benthall et al. (2021, S. 6) die Diskussionen um die Veräußerbarkeit der Daten wieder. So sei geistiges Eigentum veräußerbar, die Privatsphäre jedoch nicht. Daher würden Rechte an geistigem Eigentum von personenbezogenen Daten zu einem Widerspruch innerhalb der liberalen Rechtstheorie führen, zumindest in ihren menschenwürdigen Formulierungen. Andere hätten sich dafür ausgesprochen, die Erstellung personenbezogener Daten als eine Form von Arbeit zu behandeln, die dann auf einem Markt verkauft werden könnte. Kritikpunkt an den Datenmärkten und den daraus resultierenden formalisierten Eigentumsrechten an Daten sei, dass die politische Ökonomie der Datenverarbeitung sich wohl kaum ändern wird. Eigentumsrechte würden die Dynamik eines Regulierungssystems nicht stören. Infolgedessen würden Daten bereits wie Pseudoware behandelt, die durch Geschäftsgeheimnisse, technische Infrastrukturen und Verträge geschützt seien. Dieser Umstand ermögliche die Einschließung und primitive Akkumulation von Daten durch Plattformen. Plattformen, die durch Daten Geld verdienen, würden eher Firmen als Märkten ähneln, während sie gleichzeitig die liberale Vorstellung von einer Firma verdrängen.

5.1.3.3 Code: Datenschutz

Weitere Herausforderungen und Spannungsfelder gibt es auch beim Datenschutz von personenbezogenen Daten. Laut Zuiderveen Borgesius (2020, 10 f.) fallen algorithmische Entscheidungsprozesse dann nicht in den Anwendungsbereich des Datenschutzrechts, wenn sich KI-Anwendungen als Vorhersagemodelle nicht auf identifizierbare Personen beziehen, jedoch auf eine Gruppe oder Region. Beispielhaft wäre hier Folgendes:

> “A predictive model that says ‘80% of the people living in postal code 10017 pay their bills late’ does not refer to an individual. Therefore, the model is not a personal datum. (Data protection law does apply when such a predictive model is applied to an individual)” (ebd.).

Die Vorhersage ist zwar im ersten Anschein nicht auf eine Person rückführbar, jedoch kritisch, wenn aufgrund der Postleitzahl bspw. eine Person indirekte Diskriminierung erfährt, aufgrund der Zahlungsmoral der Nachbarschaft.

Dies hat zur Folge, dass das Spannungsverhältnis zwischen der Einhaltung des Datenschutzrechts und die Erhebung personenbezogener Daten bestehen bleibt, da laut dem Autor unklar sei, ob sich die Einhaltung des Datenschutzrechts mit der Einführung der Datenschutz-Grundverordnung (DSVGO) verbessert hätte. Seitdem wurden den Datenschutzbehörden in der EU neue Befugnisse eingeräumt, aber es sei noch zu früh, um etwas über die Auswirkungen dieser Befugnisse zu sagen (ebd.).

5.1.3.4 Code: Daten Labeling

Im Wertstrom der Daten benötigen unstrukturierte Rohdaten, nach der Datenerhebung eine Aufbereitung, die einem Produktionsprozess nahekommt. Vredenburgh (2021, S. 10) thematisiert die hyperspezialisierte Aufteilung der Datenaufbereitung. Diese sei in Mikroaufgaben aufgeteilt, die von Menschen, den sog. Crowdworkern (► vgl. Kap. 3.1.2), verrichtet werden. Der Autorin nach gibt die Ausführung von Mikroaufgaben den Arbeitenden oft keinen Zugang zu Erkenntnissen über die relevanten normativen Eigenschaften der größeren Aufgabe, z. B. ob Bilder von bestimmten sozialen Gruppen oder geopolitischen Regionen in der Datenbank unangemessen überrepräsentiert seien. Algorithmusgestützte Hyperspezialisierung sei eine epistemische Barriere am Arbeitsplatz. Sie hindere Einzelne daran, zu verstehen, wozu die Tätigkeit dient, und behindere die Entwicklung einer angemessenen praktischen Orientierung.

Neben der Hyperspezialisierung der Aufgaben beim Datenkennzeichnen, sind laut Jo et al. (2020, S. 5) feste Bezeichnungen vorgegeben, die den Crowdworkern zur Auswahl stehen. So seien sie gezwungen aus den begrenzten Optionen, die von Forschenden vorgegeben sind, eine Auswahl für die Bildkennzeichnung zu treffen.

Über die Einbeziehung aller Gruppen in die Datensätze hinaus müssten laut Bondi et al. (2021, S. 432) die Daten stark differenziert werden, um Diskrepanzen aufzudecken. Hierzu wird das Beispiel während der COVID-19-Pandemie im Dezember 2020 aufgezeigt. Das Bureau of Labor Statistics hat einen Verlust von insgesamt 140.000 Arbeitsplätzen gemeldet. Bei der differenzierten Betrachtung der Daten hat sich ergeben, dass die Verluste ausschließlich Frauen betrafen: Frauen verloren 156.000 Arbeitsplätze, während Männer 16.000 hinzugewannen, und die Arbeitslosigkeit war bei Schwarzen, lateinamerikanischen und asiatischen Frauen am größten. Ohne die Lebensumstände aller Menschen zu berücksichtigen, die von solchen Systemen betroffen sind, sei es schwierig zu behaupten, dass diese Technologien dem sozialen Wohl dienen.

5.1.4 Unterkategorie – Datengrundlage

Die differenzierte Betrachtungsweise auf Daten thematisieren auch Barocas et al. (2021, S. 371). Ihnen zufolge gibt es viele verschiedene Personengruppen, für die KI-Anwendungen eine schlechte Leistung erbringen. Darunter fallen Gruppen, die auf demografischen, soziokulturellen, verhaltensbezogenen und physischen Faktoren basieren. So hätten bspw. ethnische Zugehörigkeit, Geschlecht, Alter, Gesichtsbehaarung, Frisur, Brille, Gesichtsausdruck, Körperhaltung und Hautfarbe nachweislich einen Einfluss auf die Leistung von gesichtsbasierten KI-Systemen. Oft würde es sich dabei um Gruppen handeln, die bereits benachteiligt sind.

Für Galdon Clavell et al. (2020, S. 266) sei der Mangel an Standardisierung der Grund für die Verzerrungsprobleme in KI-Anwendungen. Daher sei das Rekonstruieren von Ursachen schwierig, wenn ein System bereits in Betrieb ist. Infolgedessen fordern sie, dass wirksame methodische Instrumente integriert werden müssen, um die Trainingsdaten im Hinblick auf die demografischen Merkmale einer Stichprobengruppe zu bewerten.

5.1.4.1 Code: Ausschluss von sensiblen Daten

Laut Sharma et al. (2020, S. 359) könnte eine einfache Pre-Processing-Technik die Verzerrungsprobleme durch „fairness through unawareness“ (► vgl. Kap. 3.2.4) lösen, bei der ein geschütztes Merkmal beim Training eines Modells nicht berücksichtigt wird, um Fairness in einer KI-Anwendung zu gewährleisten. Allerdings räumen die Publizierenden ein, dass dieser Ansatz immer noch zu Verzerrungen durch andere Merkmale führen könnte, die mit dem geschützten Attribut korrelieren. Denn wie im Kapitel 3.2.4 aufgezeigt, werden Proxies für die Vorhersage von KI-Anwendungen verwendet.

Bei „fairness through unawareness“ wäre laut Galdon Clavell et al. (2020, S. 266) bereits nachgewiesen, dass das Fehlen von Daten über einen sozialen Identifikator - ein Attribut oder geschütztes Merkmal, welches Informationen über benachteiligte Gruppen (*Race*, Geschlecht, Religion usw.) enthält - zu Verzerrungen führen. Die Publizierenden führen aus, dass diese Form der Verzerrung als ‚color blindness‘ definiert wird. Aus methodischer Sicht könne die Entwicklung einer KI-Anwendung mit color blindness die Opazität der algorithmischen Verarbeitung erhöhen und damit den Anspruch der Transparenz verwirken, da die Identifizierungsmechanismen von Verzerrungen eingeschränkt werden. Während des Lernprozesses sind ML in der Lage (► vgl. Kap. 3.2.4), die geschützten Kategorien aus den Daten abzuleiten, indem sie Proxies verwenden, die in anderen Variablen eingebettet sind, womit sie so indirekt über sensible Attribute lernen und sie in den Entscheidungsprozess einbeziehen.

Eine KI-Anwendung mit color blindness, die die Kategorie ‚*Race*' nicht berücksichtigt, würde das fundamentale Verbot der ‚*weißen* Vorherrschaft' (► vgl. Kapitel 3.3 Kelly 2021, S. 39) unwirksam machen. Laut Leavy et al. (2021, S. 698) ist die Kategorie ‚*Race*' ein integraler Bestandteil vieler Systeme der modernen Gesellschaft, z. B. des Rechts und der Strafverfolgung, der Bildung, des Gesundheitswesens, der Beschäftigung usw. Die Kategorie ‚*Race*' dient zur Identifizierung des ‚Unterdrückungsapparats' sowie den institutionellen Umgang mit dieser Kategorie. Für die Kuration von Daten sei es wichtig, dass das akademische und fachliche Wissen von Rassismus kritischen Theoretiker*innen miteinbezogen wird sowie des situierten Wissens derjenigen, die die rassistischen Auswirkungen in ihrem Alltag erleben. Der Import dieser Ideen erfordere für die Datenkuration eine aktive antirassistische Haltung. Infolgedessen bedeutet die Annahme eines ‚color blindness' Ansatzes, dass Rassismus eher verdeckt als aufdeckt wird, weil Zahlen nicht neutral und Kategorien weder natürlich noch gegeben seien sowie Daten nicht für sich selbst sprechen könnten und quantitative Daten eine aktive Rolle beim Aufzeigen und Bekämpfen von Rassismus spielen sollten.

5.1.4.2 Code: Erziehungspraktiken

Nach Croeser et al. (2019, S. 425) hätten Verzerrungen in KI-Anwendungen eine Ähnlichkeit mit der Erziehung von Kindern. Die KI-Entwicklung durch die Brille radikaler Erziehungspraktiken zu sehen, würde das Verständnis über die Trainingsdaten verbessern, da sie in der realen Welt gesammelt werden und genauso fehlerbehaftet sind wie die Welt selbst. Das sei eine Perspektive, die schnell vergessen wird, wenn der eigene Beruf und die eigene Identität mit Privilegien und Sicherheit verknüpft seien. Vor allem, wenn Datensätze von Orten gesammelt werden, die die Forschenden nicht verstehen, sollten sie daran denken, dass sie möglicherweise Strukturen widerspiegeln, von denen sie nicht möchten, dass ihre Kinder oder KI-Modelle ohne Sorgfalt und Anleitung lernen. So würden zahlreiche Beispiele von KI-Anwendungen zeigen, wie verschiedene Formen von vorurteilsbehaftetem oder antisozialem Verhalten sie nachahmen oder widerspiegeln, die Parallelen zur menschlichen Erziehung von Kindern hätten. Es würden Trainingsdaten existieren, die strategische Täuschung, Ignoranz, kognitive Dissonanz, Gewalt und eine Reihe anderer nicht hilfreicher oder destruktiver menschlicher Tendenzen lehren würden.

Hiermit ist die Untersuchung für mögliche Ursachen zu pre-existing Bias in Daten abgeschlossen. Als Nächstes werden in der Oberkategorie Outcome Lösungsansätze dargestellt und interpretierend diskutiert.

5.2 Oberkategorie – Outcome

In der Oberkategorie Outcome werden zum einen die ausgewählten Publikationen auf die Fragestellung analysiert, was in der wissenschaftlichen Literatur zur Behebung von pre-existing Bias in Daten bekannt ist. Zum anderen welche Konzepte oder Modelle in den ausgewählten Publikationen vorgeschlagen werden, die pre-existing Bias in Daten beheben.

Es sei in wissenschaftlich orientierten Kreisen laut Bondi et al. (2021, S. 427) besonders attraktiv zu behaupten, dass eine moralische Entscheidung einfach durch die Maximierung einer Zielfunktion von KI-Anwendungen getroffen werden könne. Deutlich wird durch diese Behauptung, dass die Freiheiten der atomistischen Individualität in den Hintergrund gedrängt werden, während die Aufmerksamkeit auf kollektive Interessen gelenkt wird (Benthall et al. 2021, S. 10), wenn gemeinwohlorientierte KI-Anwendungen alle Lebenswelten fair berücksichtigen soll. Wenn das Datenschutzrecht und seine Entsprechungen auf individueller Autonomie und Kontrolle zum Schutz von Würde, Persönlichkeit und Selbstdarstellung beruhen, welche Zwecke könnten dann eine sinnvolle kollektive Autonomie in der Datenwirtschaft darstellen (ebd.). Zu der kollektiven Autonomie kommt noch deren Interessen hinzu. Die Durchsetzung der kollektiven Interessen könnte durch Institutionen beschleunigt werden, wenn sie ein Rahmenwerk zur Einbindung der Gesellschaft in die Forschung festlegen und die Anreize für KI-Forschende erhöhen, die betroffene Gemeinschaften sinnvoll einbinden, um gleichzeitig wirkungsvollere Forschung zu betreiben (Bondi et al. 2021, 427 ff.).

Erste Bestrebungen gibt es seit 2017, die in Form von Konferenzen stattfinden. Angesichts der zunehmenden Verbreitung von KI und ihrer Auswirkungen auf die Gesellschaft wurde die Artificial Intelligence, Ethics and Society (AIES) ins Leben gerufen, um Raum für die Auseinandersetzung mit diesen Themen zu bieten (Acuna et al. 2021, S. 308). Allerdings sind Big-Tech Unternehmen tendenziell stärker in AIES-Konferenzen vertreten (ebd.). als bspw. die Zivilgesellschaft.

5.2.1 Unterkategorie – Datenqualität

Um die Datenqualität zu verbessern, schlagen Alvero et al. (2020, S. 205) Data Auditing vor. Entweder allein oder als Teil eines ergänzenden Rahmenwerks zur algorithmischen Prüfung. Der Kriterienkatalog AIC4 bspw. beinhaltet keine Auditierungen und empfiehlt Daten aus vertrauenswürdigen Quellen zu verwenden, die korrekt kommentiert und angemessen geschützt sind und in eigener Verantwortung der Unternehmen liegen (BSI 2021a, S. 9) (► vgl. Kap. 3.1.2).

Wohingegen Sharma et al. (2020, S. 359) die Technik der Datenerweiterung eines Datensatzes vorschlagen, der in der realen Welt erhoben wurde. Der Vorschlag

ist, ungleich verteilte Attribute mit synthetischen Datenpunkten auszugleichen, um einen ‚Idealwelt-Datensatz' zu konstruieren. Dieser neue Datensatz enthält bspw. nun eine gleiche Anzahl von Männern und Frauen, und die Kennzeichnung hängt nicht mehr vom Geschlecht ab, wodurch die Verzerrung des Modells, mit dem dieser Gesamtdatensatz trainiert wird, möglicherweise beseitigt wird. Der Ansatz kann nur dann funktionieren, wenn nur heterogene soziale Gruppen betrachtet werden, bspw. nur Frauen und Männer einer Mehrheitsgesellschaft, die keine Überschneidungen von anderen geschützten Merkmalen aufweisen. Andernfalls kann diese Vorgehensweise die Intersektionalität (►vgl. Kap. 3.2.2) von sozialen Gruppen übersehen.

5.2.1.1 Code: Diversität

Um KI-Anwendungen fairer zu gestalten, sehen Kuhlman et al. (2020, S. 5) Verbesserungspotentiale in der Vernetzung von Hochschuleinrichtungen, Einbeziehung von Interessenvertretenden aus verschiedenen Gemeinschaften in den Forschungsprozess und Förderung von diversen Gruppen künftiger Führungskräfte innerhalb der KI-Forschungsaktivitäten. Laut den Publizierenden würde eine größere Vielfalt am Tisch zwar nicht automatisch mehr Gerechtigkeit bedeuten aber durch diverse Teammitglieder würde die Belegschaft auf individueller und gemeinschaftlicher Ebene unterschiedliche Werte und Hintergründe kennenlernen.

Aus der Perspektive der Belegschaft kann Vielfalt sicherlich vorteilhaft sein, um ihre Horizonte zu erweitern. Für die, ‚die in unerwünschter Weise anders' (Scherr 2010, S. 43) sind, kann es mitunter ermüdend sein, ständig um Gerechtigkeit kämpfen zu müssen oder ihre Andersartigkeit erklären zu müssen. Daher darf der Platz am Tisch keine Alibifunktion haben, sondern volle Teilhabe ermöglichen und eine Streitkultur etablieren, die in konsensfähige Debatten und Diskussionen münden.

5.2.1.2 Code: Inklusion

Oft werden Diversität und Inklusion synonym verwendet, um Gleichheit zu suggerieren, wobei es für Chi et al. (Chi et al. 2021, 449 ff.) klare Unterschiede zwischen diesen Begriffen gibt. Während Diversität die Repräsentation verschiedener Gruppenzugehörigkeiten und deren kultureller Unterschied meint, ist Inklusion die Einbeziehung dieser Gruppen in wichtige Entscheidungsprozesse und mit Gleichheit wird die Abwesenheit von struktureller Diskriminierung gefördert. Ferner sei Inklusion entscheidend für die Demokratisierung der KI-Anwendungen, die Förderung von Kund*innenvertrauen und die Gewinnung besserer Geschäftserkenntnisse.

Auch Leavy et al. (2021, 699 f.) sind für die Demokratisierung von Daten. Durch die Einbeziehung aller Perspektiven, insbesondere derjenigen, die am stärksten von Diskriminierung betroffen seien, könne die Entwicklung von Technologien

vorangetrieben werden. Zudem würde durch das Einbeziehen von marginalisiertem Wissen die Datenqualität sich verbessern. Außerdem würde dieses Wissen eine schnellere Lösungen zur Erreichung kollektiver Ziele bieten (Bondi et al. 2021, S. 432).

5.2.1.3 Code: Personal

Gilbert et al. (Gilbert et al. 2019, S. 66) empfehlen, das KI-Ingenieur*innen hinsichtlich eines gesunden Arbeitsumfelds geschult werden sollten. Dazu gehört der Zugang zur Rechtsberatung und die Kultivierung von emotionaler Intelligenz (engl. Soft Skills), um Unstimmigkeiten besser verarbeiten zu können. Zu den Schulungsmaßnahmen gehört des Weiteren ein allgemeines Verständnis über den Einsatz von soziotechnischen KI-Anwendungen.

5.2.1.4 Code: Interdisziplinäre KI-Entwicklung

Laut Bondi et al. (2021, S. 433) sei ein gutes Beispiel das Center for Analytical Approaches to Social Innovation (CAASI) an der University of Pittsburgh, das interdisziplinäre Teams aus den Bereichen Politik, Informatik und Sozialwissenschaft zusammenbringt. Insbesondere die Zusammenarbeit mit Sozialwissenschaften, die sich mit der Bewertung von Mechanismen struktureller Diskriminierung befassen, sei dieser Umstand für das Fachgebiet von großem Nutzen (Kuhlman et al. 2020, S. 6). Das Einbringen der verschiedenen Fachdisziplinen in die KI-Entwicklung ist der erste Schritt den Menschen hinter den mathematischen Konstrukten zu sehen. Angesichts der weitreichenden Veränderungen, die die neuen Technologien mit sich bringen, sei eine gemeinsame Anstrengung von Sozialwissenschaften und KI-Forscherenden erforderlich (Cruz Cortés et al. 2020, S. 235). Allerdings sollten KI-Forschende laut den Publizierenden diese wichtigen Szenarien nicht vollständig auf Kolleg*innen aus der Sozialwissenschaft oder dem Non-Profit-Bereich abwälzen (Bondi et al. 2021, S. 433). Das interdisziplinäre Teilgebiet könnte sich auf die Erfassung, den Austausch, die Kennzeichnung, die ethische Überwachung und die Aufzeichnungsprozesse von Daten konzentrieren (Jo et al. 2020, S. 2). Ein Interventionssystem aus mehreren Personen kann systematischer einen Datensatz zusammenstellen als eine einzelne ML-Ingenieur*in (ebd.). Weitere Maßnahmen wären Kontrollinstanzen innerhalb und zwischen den Institutionen zu installieren, um die Datensammelnden zu kontrollieren (ebd.). In der KI-Gemeinschaft gibt es derzeit keine dieser Sicherheitsvorkehrungen (ebd.).

Leavy et al. (2021, S. 701) fordern, wenn sprachbasierte Trainingsdatensätze im Hinblick auf Rassismus/Misogynie und andere Hassdiskurse ausgewertet oder kuratiert werden sollen, muss die Einbeziehung von Teammitgliedern mit Fachwissen im Bereich der Gender und Rassismus kritische Theorien als notwendige Voraussetzung

angesehen werden. Angesichts der Veränderungen und der Kontextgebundenheit rassistischer und geschlechtsspezifischer toxischer Sprache würden situiertes Wissen einen erheblichen Mehrwert für theoriegestützte Ansätze darstellen. Ferner sollten im Design Thinking Ansatz Menschen beteiligt sein, die Rassismus erleben oder produzieren sowie voll integriert und angemessen entschädigt werden. Die Auswertung und Kuration von Trainingsdaten für KI-Anwendungen erfordert infolgedessen die Bildung eines multidisziplinären Teams mit speziellen Fähigkeiten und den Einsatz verschiedener Formen der Wissensproduktion.

5.2.1.5 Code: Feministische Theorien

Angesichts der offensichtlichen Bedrohung der sozialen Gerechtigkeit und Menschenrechte durch lernende KI-Anwendungen, läge es auf der Hand, dass die Methoden für die KI-Datenkuration grundlegend geändert werden müssen (Leavy et al. 2021, S. 701). Die Einbeziehung von kritischen Theorien über Intersektionalität, Feminismus und Rassismus könne dazu dienen, den Rahmen für die Fairness Evaluierung bzgl. der Repräsentation von Gruppen in Daten zu erweitern. Darüber hinaus, so die Publizierenden, lenkt die Einbeziehung solcher Perspektiven die Aufmerksamkeit weg von der Suche nach technischen Verzerrungen hin zu einer umfassenderen Untersuchung der Machtstrukturen und Ideologien von Geschlecht, *Race* und sozialer Klasse, die in den Daten enthalten sein können (ebd.). Solche theoretischen Rahmungen können zum Gestaltungsprozess der Datenlieferkette beitragen, die eine intersektionale und feministische Repräsentation widerspiegelt (ebd.).

5.2.1.6 Code: Datenschutz

Gleichstellungs- und Datenschutzbehörden sollten über ausreichende Ermittlungs- und Durchsetzungsbefugnisse verfügen und mit angemessenen Mitteln ausgestattet werden, um bspw. technisches Fachwissen einzustellen (Zuiderveen Borgesius 2020, S. 12).

5.2.1.7 Code: Partizipation

Laut Acuna et al. (2021, S. 308) könnten KI-Ethikkonferenzen Forschende oder die Öffentlichkeit aktiv zur Teilnahme an Diskussionen einladen. Wenn Fragen der Voreingenommenheit in KI-Anwendungen auf sinnvolle Weise angegangen werden sollen, legen Liberalismus und soziale Gerechtigkeit nahe, so die Publizierenden, dass die Meinungen derjenigen angehört werden sollten, die am meisten von der Technologie betroffen sind. Beim Participatory Design bzw. Co-Design und Design Justice Ansatz werden nicht nur die Meinungen angehört, sondern alle Beteiligten und Betroffenen in die KI-Einwicklung einbezogen (Leavy et al. 2021, S. 698). Diese Design Ansätze zielen

darauf ab, strukturelle Diskriminierung durch Designpraktiken zu beseitigen, indem sie das Wissen und die Erfahrungen von Betroffenen einbeziehen und priorisieren (ebd.). Partizipative Ansätze seien oft am erfolgreichsten, wenn KI-Forschende ihre gewohnten Sphären verlassen und sich in die Bereiche begeben, auf die sich ihre Projekte auswirken (Bondi et al. 2021, S. 430). Im Designprozess existieren keine hierarchischen Ebenen, während Designer*innen oder KI-Entwickelnde als Vermittelnde fungieren, übernehmen Betroffene, aufgrund ihrer Erfahrungen, die Rolle von Fachexpert*innen (ebd.). Kollaborative Ansätze können die Entwicklung von kultursensiblen KI-Anwendungen unterstützen, die einen proaktiven, inklusiven Ansatz verfolgen und Interkulturalität, Innovationen und technosozialen Aktivismus berücksichtigen (Kuhlman et al. 2020, S. 6). Diese Ansätze können in allen Phasen der Datenaufbereitung wie Sammlung, Analyse und Interpretation von Daten angewendet werden (ebd.). Um werteorientierte Technologie zu entwickeln, ist das Value Sensitive Design ein weiterer möglicher Ansatz, die Unterschiedlichkeit der Nutzenden zu berücksichtigen. Der Sinn eines Value Sensitive Designs besteht nicht darin, ein Set von akzeptablen Werten vorzudiktieren, sondern vielmehr darin, die Mittel bereitzustellen, um das ‚Richtige' in einem bestimmten Kontext zu bestimmen (Schelenz et al. 2021, S. 907).

5.2.1.8 Code: Aktivistische Haltung

Werte, die sich durch den Buchdruck wesentlich verändert haben, stellen nach Posada et al (2021, S. 870) eine Möglichkeit dar, Analogien zwischen dem Buchdruck in der Frühzeit und den jetzigen Diskussionen rund um die rechtlichen Rahmenbedingungen von KI-Anwendungen herzustellen. Sie fragen sich: Welche Regeln wurden in der Welt des Buchdrucks möglich, die in der Welt des gesprochenen Wortes unmöglich waren? Welche Regeln sind in einer Welt der digitalen Daten unmöglich geworden, die in der Welt des Buchdrucks möglich waren? In ihrer Analogie stellen sie fest, dass es hinsichtlich der Menschenrechte, Meinungsfreiheit, Recht auf Privatsphäre und Einwilligung Ähnlichkeiten gibt. Daher könne die Gesellschaft nicht davon ausgehen, dass die Regeln von vor zwanzig Jahren jetzt oder in zwanzig Jahren noch gelten werden. Stattdessen soll die Gesellschaft aus den Auswirkungen des Buchdrucks in der Frühzeit lernen und sich fragen, wer in einer digitalen Welt die Macht hat, warum, und welche konstitutiven Veränderungen von der Technologie zu erwarten seien.

Aktivistische Haltung fordern auch Leavy et al. (2021, 698 ff.) beim Kuratieren von Daten für KI-Anwendungen, die im Einklang mit feministischen Grundsätzen und der Rassismus kritischen Theorie stehen sollen. Die feministische Wissenschaftstheorie sei sich bewusst, dass Entscheidungen, die bei der Darstellung von Wissen getroffen werden, eine Rolle bei der Entwicklung gesellschaftlicher Konzepte spielen. Daten sollen

daher mit einer antirassistischen Haltung bewertet werden, in dem Bewusstsein, dass die Kategorie *Race* durch Daten gemacht und legitimiert wird. Dieses Bewusstsein würde dazu dienen, Rassismus in der Gesellschaft zu bekämpfen.

5.2.2 Unterkategorie – Datensammlung

Für die Unterkategorie Datensammlung wurde durch die qualitative Inhaltsanalyse in den Publikationen keine Hinweise zur Behebung gefunden. Daher entfällt diese Unterkategorie in der SLR.

5.2.3 Unterkategorie – Datenerhebung

Um Kategorisierungen von sensiblen Merkmalen zu begreifen, ist es laut Kshirsagar et al. (2021, S. 667) wichtig, bei der Modellierung von gemeinwohlorientierten KI-Anwendungen, die zugehörigen Metadaten, den Erfassungsprozess und etwaige Sicherheits- oder Datenschutzbedenken von Datensätzen zu verstehen. Dies könnte bereits in der direkten Datenerhebung mit den Betroffenen geschehen, um die Akzeptanz der KI-Anwendungen zu erhöhen (Barocas et al. 2021, S. 375). Schließlich biete dieser Ansatz auch die vollständige Kontrolle über die Lizenzierung und würde es vereinfachen, die Zustimmung der Betroffenen einzuholen und sie für die Bereitstellung ihrer Daten fair zu entschädigen (ebd.). Bei diesem Ansatz sollten auch die Möglichkeiten des Widerrufsrechts etabliert werden. Denn Lebensumstände ändern sich und das Recht auf Vergessenheit in der Datensphäre (Bundesdatenschutzgesetz § 35) sollte hier durch die Veräußerung der personenbezogenen Daten nicht aufgehoben werden.

Über dies würde die Forderung von Gilbert et al. (2019, S. 65) den vorrangegangenen Vorschlag ergänzen, indem Daten nicht nur kontextspezifisch erhoben, sondern auch dokumentiert werden, um Transparenz und Rechenschaftspflicht in der Entwicklungspipeline zu verankern. Zudem gehört auch, dass Daten nicht nur explizit mit Kommentaren versehen werden, um implizite Annahmen über die Daten vorzubeugen, sondern auch die Veröffentlichung der erhobenen Daten, damit die dahinterstehenden Ideologien von denjenigen problematisiert werden können, die am stärksten von ihrer Kennzeichnung oder Klassifizierung betroffen sind (ebd.). Kshirsagar et al. (2021, S. 667) betonen, dass in mehreren gesellschaftlich wichtigen Bereichen die Bezeichnungen in Datensätzen unter subjektiven Kommentaren leiden. Solche Situationen sollten im Vorfeld erkannt werden, um Inkonsistenzen in der KI-Entwicklung zu vermeiden (ebd.). Damit wird deutlich, dass Datenerhebung kein statischer Vorgang ist, sondern ein Produkt des Kompromisses, welches ständige Reflexion und Überlegung benötigt (Gilbert et al. 2019, S. 65). Dieser Kompromiss erfordert auch, dass bei einer diversitätsbewussten Datenerhebung, die lokalen Gewohnheiten berücksichtigt

werden sollten, die sich in geografischen und kulturellen Regionen unterscheiden können (Schelenz et al. 2021, S. 912). Zu dem Kompromiss gehört auch, dass bei kontinuierlichen Aktualisierungen, der Mensch sowohl in den Prozess der Datenerfassung sowie Kennzeichnung der Daten, als auch bei der KI-Entwicklung und Modellanpassung eingebunden werden sollte (Kshirsagar et al. 2021, S. 669).

5.2.3.1 Code: Professionalisierung der Datenerhebung

Jo et al. (2020, S. 8) machen den Vorschlag die Datenerhebung zu professionalisieren, um die Anreize zur Einhaltung ethischer Richtlinien zu erhöhen. Vergleichbar mit dem Beruf einer Archivar*in, könnte die Arbeit einer Datenerhebenden, ausgestattet mit dem Wissen der Ethikkodizes, nicht nur Daten sammeln, sondern auch Beratungsfunktionen beim Einsatz der Daten übernehmen. In diesem Sinne könnte die Einrichtung einer institutionsübergreifenden Organisation, dazu beitragen, dass ethische Grundsätze den gewinnorientierten Motiven von Unternehmen widerstehen. Infolgedessen könnten die Durchsetzungsstrategien und Lizenzbedingungen von Archiv- und Bibliothekswissenschaften als Vorbild für die KI-Community dienen.

5.2.3.2 Code: Monetarisierung

Ein alternatives Modell könnte laut Benthall et al. (2021, 8 ff.) eine Kooperationsgemeinschaft sein. Als Mitglieder eines intelligenten sozialen Systems entwickeln ihre eigenen Such- und Empfehlungsprioritäten und bestimmen selbst über ihre Daten, um Einnahmen zu generieren, die anteilig an die Mitglieder ausgezahlt werden (ebd.). Hierbei sollte jedoch der Gewinn nicht der einzige Zweck sein. Vielmehr wäre der Zweck des Systems die Selbstkontrolle und Selbstverwaltung (ebd.). Vor diesem Hintergrund ist die Befähigung einer Gemeinschaft unabdingbar, die ihre eigenen Daten als Gemeingut selbst verwalten (Schelenz et al. 2021, S. 913). Eine Auseinandersetzung mit dem Bereich des Organisationsrechts sei nach Benthall et al. (2021, 8 ff.) für KI-Ethiker*innen eventuell inspirierend. Dem Kollektivverständnis ähnlich, sieht der Richtlinienentwurf der indischen Regierung vor, dass die Daten der Inder*innen als eine kollektive Ressource bzw. als ein nationales Gut verstanden werden, welches die Regierung treuhänderisch verwalten möchte (Kak 2020, S. 309). So wie Ausländer*innen in Indien keinen gleichberechtigten Zugang zu Kohleminen oder Fernmeldeleitungen erhalten, argumentiert der Entwurf der Politik, dass indische Bürger*innen und Unternehmen bei den wirtschaftlichen Vorteilen aus der Monetarisierung von Daten Vorrang haben müssen.

5.2.3.3 Code: Intention

Bei der Datenerhebung ist laut Jo et al. (2020, 4 f.) als Erstes zu fragen, was sind die Motive der Datenerhebung und welcher Intention dienen sie. Insbesondere dann,

wenn mit Datensätzen, die aus dem Internet kommen, KI-Anwendungen trainiert werden. Den Publizierenden nach sollte diese Vorgehensweise mit einer Interventionsebene versehen werden.

Die Interventionsebene ist in der DSGVO festgeschrieben, die Galdon Clavell et al. (2020, S. 266) als Datenminimierung in Datensätzen versteht. Ihnen nach sollten alle Kategorien von sensiblen Merkmalen in Daten entfernt werden, wenn sie für die Aufgabenstellung nicht erforderlich sind. Daher müsse die Erhebung von personenbezogenen Daten auf ein Mindestmaß beschränkt werden, um die Ziele bei der Datenverarbeitung zu erreichen.

5.2.4 Unterkategorie – Datengrundlage

Die derzeitigen Strategien zur Verringerung von Verzerrungen fassen Tommasi et al. (2021, 1 f.) in drei Hauptgruppen zusammen. Im Pre-Processing-Verfahren werden aus den Datensätzen ‚unfaire' Daten entfernt bzw. durch synthetische Datenpunkte angeglichen und im In-Processing-Verfahren werden während des Trainings eines KI-Modells Parametermodifizierungen vorgenommen. Dann gibt es noch das Post-Processing-Verfahren, bei dem eine Output-Korrektur vorgenommen wird.

Kshirsagar et al. (2021, S. 668) legen noch nahe, dass eine sorgfältige Überlegung hinsichtlich der Datenaufteilung in Trainings- und Testdatensätze notwendig ist, um die Generalisierungsfähigkeit des KI-Modells für unbekannte Dateninstanzen messen zu können. Um das Verzerrungsrisiko in KI-Anwendungen zu mindern und eine Fairness-Zertifizierung für Daten zu etablieren, wäre laut Tommasi et al. (2021, S. 2) hilfreich, wenn alle Faktoren über die Daten in einer Liste dokumentiert werden. Insbesondere sollen die geschützten Attribute in jeder Datensammlung genau definiert und deren kategoriale Zuordnungen festgehalten werden sowie statistische Hypothesentests für jedes Attribut durchgeführt und Bewertungsmetriken festgelegt werden, die für die Zertifizierung als sicher gelten (ebd.). Laut Barocas et al. (2021, S. 370) sei der Entwurf von Evaluierungen für KI-Anwendungen, die sich auf eine bestimmte Gruppe von Menschen konzentrieren, in der Regel einfacher als einen Entwurf zu entwickeln, der allgemeingültig sein soll. Deshalb sei eine konfirmatorische Evaluierung (eine Hypothesen bestätigende oder widerlegende Untersuchung) am ehesten möglich, wenn die Systemleistung für eine kleine Anzahl besonders auffälliger Gruppen in Szenarien bewertet und untersucht wird, in denen es nur wenige zusätzliche Faktoren gibt, die die Systemleistung beeinflussen können (ebd.).

Die aktuellen Best Practices sind laut Bryant et al. (2019, S. 382) die angemessene Repräsentativität von Populationen in Trainingsdatensätzen sicherstellen, die Validität von Datensätzen in der Trainingsphase bewerten und

Evaluierungen durch unabhängige Stellen für Erkennungs-, Klassifizierungs- und Empfehlungssysteme. Mit diesen Best Practices könne die KI-Technologie schrittweise weiterentwickelt werden (ebd.). Diese Best Practices sind entstanden, weil KI-Anwendungen ungerechte Prognosen getroffen haben. Sie sind erst auffällig geworden, als sie Menschenleben bewerteten oder bei Schwarzen Bevölkerungsgruppen schlechter funktionierten als bei *weißen*. ProPublica evaluierte das COMPAS-Rückfallprognosesystem von Northpointe (Angwin et al. 2016, o.S.). Forschende evaluierten die Spracherkennungssysteme von Amazon, Apple, Google, IBM und Microsoft (Koenecke et al. 2020, S. 7684). KI-Gesichtserkennungsprogramme von IBM, Microsoft, and Face++ wurden von Forschenden evaluiert, die die Gender Shades Studie durchgeführt haben (Buolamwini 2019, S. 2). Diese drei Vorfälle sind ans Tageslicht gekommen, weil die Funktionsweise der KI-Anwendung anhaltende strukturelle Diskriminierung aufgewiesen hat, deren Überprüfung weder von staatlichen Seiten kam noch von den Unternehmen selbst, sondern von NGO‘s oder Forschende einer Universität. Demnach wäre die Überlegung wert, dass solche Einrichtungen die Evaluierung in Form von Audits und Zertifizierungen von KI-Anwendungen als Unabhängige durchführen und entsprechende Ressourcen erhalten. Bei Evaluierungen in diesem Format benötigen die Durchführenden Ressourcen, Fachwissen und technische Infrastruktur (Jo et al. 2020, S. 6), die ihnen von staatlicher Seite zur Verfügung gestellt werden könnten. Dafür könnten Privatunternehmen, die KI-Anwendungen entwickeln und als soziotechnisches System einsetzen wollen, Steuern an den Staat zahlen.

5.2.4.1 Code: Disaggregierte Evaluierungen von Datensätzen

Um Datensätze von KI-Anwendungen zu untersuchen, die im soziotechnischen Kontext eingesetzt werden, wird der Ansatz der disaggregierten Evaluierung angewendet. In der Vergangenheit wurden mit dieser Methode auffällige KI-Anwendungen identifiziert, die bestimmte soziale Gruppen diskriminiert haben (s.o.). Disaggregierte Evaluierungen von Datensätzen sind Methoden, bei denen statistische Daten nach bestimmten sozialen Gruppen in unterschiedlichen Einzelgrößen aufgeschlüsselt werden, um eine getrennte Bewertung für verschiedene Personengruppen durch KI-Anwendungen durchzuführen. Die zu untersuchenden Datensätze werden hierbei entweder um die unterrepräsentierten Gruppen erweitert, oder es wird ein ganz neuer Datensatz erstellt, oder es werden Datensätze verwendet, die zu Forschungszwecken veröffentlicht wurden, um die Funktionsweise von KI-Anwendungen zu überprüfen (Barocas et al. 2021, S. 374). Um Datensätze mit fehlenden Daten zu erweitern, wird die Scraping-Methode angewandt (ebd.). In diesem Fall werden Daten aus dem Internet heruntergeladen und die zu bewertenden

Datensätzen um die fehlenden Datenpunkte erweitert. Dabei ist jedoch der sorgfältige Umgang mit sensiblen Daten zur Wahrung der Privatsphäre notwendig, die zusätzliche demografische Kennzeichnungen der Daten erfordert und mit Arbeitskosten einhergehen (Jo et al. 2020, S. 1). Disaggregierte Evaluierungen benötigen zu dem vielfältige Trainingsdaten über verschiedene demografische Merkmale, um diese getrennt bewerten zu können (ebd.). Mit dieser Vorgehensweise wäre jedoch sichergestellt, dass ausreichende Daten über die unterrepräsentierten Gruppen vorhanden sind und die Population widerspiegeln, die von Interesse ist - also Personen, die in der Vergangenheit mit der KI-Anwendung in Berührung gekommen sind, oder Personen, die in der Zukunft mit der KI-Anwendung in Berührung kommen werden - sowie die Umwelt- und Verhaltensfaktoren (Barocas et al. 2021, S. 374). Bei der disaggregierten Evaluierungsmethode der Trainingsdaten stellt sich jedoch die Frage, ob sich auf soziale Konstrukte wie Rasse und Geschlecht oder auf beobachtbare Eigenschaften wie Hautfarbe, Gesichtsbehaarung und Frisur konzentriert werden sollte. Im Gegensatz zu sichtbaren Merkmalen seien *Race* und Gender keine objektiven, inhärenten Eigenschaften von Menschen, sondern konstruierte Kategorien, die aufgrund sozialer Konventionen als Grundlage für die soziale Differenzierung dienen (ebd.). Diese sozialen Konstrukte seien tief in der Gesellschaft verankert und häufig als selbstverständlich angesehen, dass sie das Verständnis der Menschen von anderen und von sich selbst strukturieren würden (ebd.). Dabei seien sie historisch und kulturell spezifisch, instabil und umstritten und oft mit ungerechten sozialen Hierarchien verbunden. Selbst wenn die Beobachtung sich auf sichtbare Eigenschaften konzentriert, so könne dies von besonderem Interesse sein, wenn sie als Proxies für soziale Konstrukte dienen (z. B. die Hautfarbe als Proxies für *Race*).

5.2.4.2 Code: Reflexive Haltung

Leavy et al. (2021, S. 700) fordern eine reflexive Haltung gegenüber Daten einzunehmen. Aus Sicht der feministischen Wissenschaftstheorie wird das in Daten enthaltene Wissen aus bestimmten gesellschaftlichen Perspektiven konstruiert und von Faktoren wie Gender, *Race*, Class und Standort beeinflusst. Der reflexive Charakter der Wissensproduktion sei entscheidend für eine ethische Datenkuration. In den meisten KI-Anwendungen seien Daten nicht statisch, sondern würden durch das Hinzufügen neuer Daten, Proxy-Variablen oder die automatische Reorganisation und Neuklassifizierung von Personen, ständig aktualisiert. Aus diesem Grund ist eine reflexive Haltung wichtig, wenn Vorschläge zu Verzerrungsminimierung gemacht werden, die technischer Natur sind. Die folgende Publikation geht davon aus, dass KI-Vorhersagen im Zusammenhang mit kriminellen Handlungen häufig mit rassistischen oder religiösen Verzerrungen behaftet seien:

> "Design: historical/external bias. **Given predictions related to criminal acts are often accused of racial or faith-based biases**, practitioners could check model performance against racial and faith groups, if these features are available from the data. If not, it could be possible to check model performance by **region, which may be acting as a proxy for *race* or religion, to assess whether high-minority-group areas** are more prone to model errors. Regarding **socioeconomic biases**, the developer could check model performance by **income** level while controlling for the ratio of claim amount to income." (Lee et al. 2021, S. 710).

Rational betrachtet trifft diese Aussage für die meisten soziotechnischen KI-Anwendungen zu, die in der Vergangenheit aufgrund ihrer Vorsagen auffällig geworden sind (s.o.). Aber diese Aussage hat m.E. eine tiefgehende Bedeutung. Es wird zwischen den Zeilen nach Scherrs Worten (2010, S. 43) die Unterscheidungspraxis zwischen ‚den Normalen' und den ‚in unerwünschter Weise anderen' verstärkt. Die ‚Normalen' (*weiß*, männlich, jung und in Gatekeeper Positionen) bekommen Best Practice Empfehlungen, wie sie die Symptome bekämpfen können, jedoch nicht die Ursachen, die sie selbst verursachen. Die problematischen Daten der ‚in unerwünschter Weise anderen' werden mit ein paar Checks weggeklickt.

Daher ist es laut Leavy et al. (2021, 698 f.) wichtig, die Perspektiven der Beteiligten zu verstehen, die in den Prozess der Bearbeitung von Trainingsdaten einfließen, einschließlich derer, die die Struktur von Datensätzen entwerfen, Daten erstellen, Kategorien für Datenkennzeichnungen entwerfen und mit Anmerkungen versehen. Die Publizierenden fordern sich zu fragen, wessen Perspektiven in den Daten kodiert sind und welche potenziellen Auswirkungen damit gefördert werden, damit KI-Anwendungen keine schädlichen sozialen Konstruktionen erlernen. Darüber hinaus würde sich die Datengrundlage verbessern, wenn multidisziplinäre Teams in Bezug auf historische Trainingsdaten eingesetzt werden, die nicht nur Trainingsdatensätze auf rassische und geschlechtsspezifische Toxizität hin untersuchen, sondern auch die Konstruktion synthetischer Datensätze mit dem Wissen und den Erfahrungen von unterworfenen und unterdrückten Gruppen erweitern (Leavy et al. 2021, S. 701). Diese Vorgehensweise würde laut den Publizierenden zu einer ‚Datenutopie' führen (ebd.). In dieser Datenutopie würden Datensätze in der Sprache Gleichheit, Freiheit und Emanzipation ‚sprechen' und KI-Anwendungen veranlassen, in diesen Sprachen zu lernen, die sich m.E. zu einer feministischen KI-Anwendung transformieren.

5.2.4.3 Code: Radikale Elternschaft

Nicht nur die Sprache, sondern die Sichtweise auf KI-Anwendungen zu ändern, indem sie als Kinder betrachtet werden, kann laut Croeser et al. (2019, 424 ff.) für das Narrativ und das Selbstverständnis der KI-Forschung ein wirksames Instrument für Veränderungen an dieser Front sein, um die kulturellen Fäden zu entwirren, die die KI-Forschung an einen Mangel an Vielfalt binden. Die überwältigende Mehrheit der KI-

Entwickelnden ist männlich, weshalb das gesamte Unterfangen ein sehr männlicher Versuch sei, der als eine Art Alleinerziehende-Elternschaft angesehen werden kann. Laut den Publizierenden würden andere Autor*innen die Mutterschaft ausdrücklich als eine Technologie der Transformation positionieren. Radikale Elternschaft erfordere die Förderung der Andersartigkeit, die in Form von unerwarteten Lebensentwürfen, Bedürfnissen, Weltanschauungen, geschlechtlichen, sexuellen oder politischen Identitäten auftreten könne. Ein Ansatz zur KI-Entwicklung in diesem Paradigma würde das Potenzial für Unterschiede schätzen.

Um den Ansatz der radikalen Elternschaft zu erweitern, könnten noch Ansätze der neueren integral evolutionären Organisationsformen in das Paradigma miteinbezogen werden. In integral evolutionären Organisationsformen ist der Mensch erst dann vollkommen, wenn er in seiner ‚Ganzheit' seinen Beruf ausüben darf (Laloux 2017, S. 86). Während traditionelle Organisationen, die von den patriarchalen Ideologien geprägt sind, die männliche Energien wie Entschlossenheit, Rationalität, Gelassenheit usw. wertschätzen, sind weibliche Energien wie Fürsorge. Empfindsamkeit, Entschleunigung usw. nicht gern gesehen (ebd.). Dabei ist zu beachten, dass jeder Mensch (egal ob Mann, Frau oder Diverse) diese Energien in sich trägt, die in der Zusammensetzung den Menschen als Ganzes ausmachen (ebd.). Um die gesellschaftlichen Machtstrukturen zu dekonstruieren könnte die Sichtweise der evolutionären Organisationsformen bei der KI-Entwicklung förderlich sein.

Hiermit ist die Untersuchung für mögliche Behebungen zu pre-existing Bias in Daten abgeschlossen. Als Nächstes werden Konzepte und Modell in der Oberkategorie Outcome dargestellt und interpretierend diskutiert.

5.2.5 Konzepte oder Modelle

Benthall et al. (2021, S. 10) berichten von digitalen Organisationsformen, wie bspw. von DAO (Distributed Autonomous Organizations). DAO sind dezentralisierte autonome Organisationen, die Blockchain-basierte KI-Einheiten nutzen. Ihre Communitymitglieder führen die Organisation gleichberechtigt (ebd.). Bei dieser Art der Organisation wird die Führungsebene durch einen Open-Source-Code ersetzt und nach dem Motto „Code is law" nach computercodierten Regeln (Smart Contracts) geführt (BTC-ECHO 2022, o.S.). DAO-Modelle würden effiziente Systeme schaffen, indem sie den Bedarf an menschlichen Eingaben reduzieren (ebd.). Das System funktioniert durch Anreize und Sanktionen (Troncoso et al. 2022, S. 7). Als Reaktion auf die individualistischen, techno-deterministischen und männlich dominierenden Führungsformen von DAO, wurde die Distributed Cooperative Organization (DisCO.coop) gegründet (Troncoso et al. 2022, S. 21). DisCO ist eine feministische und gemeinwohlorientierte Kooperative, die sich mithilfe von digitalen Werkzeugen auf

soziale und ökologische digitale Arbeit konzentriert (Troncoso et al. 2022, S. 11). Während DAO-Modelle *dezentralisiert* und *autonom* sind, ist das DisCo-Modell *verteilt* und *kooperativ*. Beide Organisationsmodelle sind selbstführend, wobei das DAO-Modell eher der traditionell konformistischen Kultur (Laloux 2015, S. 22) ähnelt, weil das Verhalten der Communitymitglieder durch Anreize und Sanktionen kontrolliert wird und unabhängig sowie autonom agiert. In traditionell konformistischen Organisationen ist die soziale Zugehörigkeit entscheidend, die mit der Haltung „wir gegen die anderen" geprägt ist (ebd.). Die Moral wird zudem in solchen Organisationsformen durch Effektivität ersetzt, die als Maßstab für die Entscheidungsfindung dient (Laloux 2015, S. 23). Das DisCo-Modell als ganzheitlicher Ansatz hingegen übernimmt das DAO-Modell und erweitert es um die sozialen Aspekte einer Gesellschaft (Troncoso et al. 2022, S. 31). Der ganzheitlicher Ansatz als sinnstiftende Formen der Zusammenarbeit, lässt sich auch in der integral evolutionären Organisationsform finden, die die Rationalität mit emotionalen, intuitiven und spirituellen Aspekten erweitert (Laloux 2015, S. 55). Ähnliche Ansätze, die gegen die Hegemonie des Silicon Valleys als Widerstand zum Überwachungskapitalismus entwickelt wurden, sind Data Co-Operativism und Data common (Kak 2020, S. 310). Die Datenmodelle werden von Stadtverwaltungen in Barcelona und Amsterdam im Rahmen des EU-Projekts DECODE (DEcentralised Citizen-owned Data Ecosystems) erprobt. Das Projekt in Amsterdam experimentiert mit einem dezentralisierten Datenökosystem in Bürger*innenhand und Barcelonas Datenmodell Decidim.barcelona konzentriert sich auf die Entwicklung von partizipativen Prozessen rund um Datenzugang und algorithmische Transparenz. Die Plattform vTaiwan ermöglicht die konsensorientierte Zusammenarbeit, die mit Expert*innen und relevanten Mitgliedern der Gesellschaft über kontroverse politische Themen diskutieren und gemeinsam Lösungsansätze entwickeln (Bondi et al. 2021, S. 431). Kollaborative Ansätze gibt es auch im Projekt Historypin, die als Community Archives, auch bekannt als Stammesarchive oder partizipatorische Archive das Sammeln von Daten oder Dokumenten ermöglichen, die im Besitz einer Gruppe sind und sich selbst repräsentieren (Jo et al. 2020, 5 f.). Community Archives sind durch die Notwendigkeit motiviert, die Stimmen der „non-elites, the grassroots, the marginalized" zu vertreten (ebd.). Das Projekt Respect als die Stimme der LGBTQ+-Gemeinschaft, sammelt positive Begrifflichkeiten, um die negativen sprachlichen Assoziationen im Zusammenhang mit geschlechtlichen Minderheiten im Internet auszugleichen (Jo et al. 2020, S. 5). Ebenso ermöglicht das MIDATA-Modell als eine Genossenschaft den Mitgliedern einen aktiven Beitrag zu medizinischen Forschungen und klinischen Studien zu leisten, indem die Mitglieder einen selektiven Zugang zu ihren persönlichen Daten gewähren und anteilig vergütet werden (Benthall et al. 2021, S. 10). Die WeNet-Plattform und -App, die als Forschungsvorhaben entwickelt wurde, umfasst ausdrücklich eine

Familie von diversitätsbewussten KI-Modellen, die die menschliche Interaktion unterstützen (ebd.). Das Projekt hat das endgültige Ziel, eine Forschungsinfrastruktur zu entwickeln, die nicht nur die gesammelten, vollständig anonymisierten Daten verwaltet und zur Verfügung stellt, sondern auch die Dokumentation der erforderlichen Prozesse, die für die gemeinsame Nutzung und (Wieder-)Verwendung der gesammelten Daten erforderlich sind, um so Transparenz und Replizierbarkeit zu ermöglichen (Schelenz et al. 2021, 906 f.). Mit Computersimulationen wie Agent Based Models (ABM) können vielfältige soziale Dynamiken eines Kollektivs erzeugt und politische Experimente durchgeführt werden (Cruz Cortés et al. 2020, S. 237). Sie unterscheiden sich von anderen KI-Modellen insofern, dass sie eine flexiblere Heterogenität in ihren Populationen zulassen, die sie durch Agenten darstellen (ebd.). So hat jeder Agent in der Computersimulationen seine eigenen Population mit bestimmten Merkmalen und handelt oft entsprechend seiner lokalen Umgebung und mit unvollkommenen Informationen (ebd.). Dieses Verhalten der ABMs sei realistischer als viele klassischen KI-Modelle, bei denen perfektes Wissen und Rationalität sowie eine homogene Population angenommen wird (ebd.). Deutlich wird bei diesen vorgestellten Projekten, dass der Top-Down Führungsstil durch kollaborative, sinnstiftende und kooperative Organisationsmodelle ersetzt werden. Im Vordergrund der vorgestellten Modelle steht die ganzheitliche Berücksichtigung der sozialen Aspekte und die Einbeziehung sowie Inklusion aller Beteiligten. Mitglieder sind ausgestattet mit Teilhabemöglichkeiten, die den feministischen Grundsätzen entsprechen. Die vorgestellten Modelle widersprechen der Datenstrategie der Bundesregierung, die Datentreuhandmodelle etablieren möchte, um zügig das „Datenteilen und -nutzen“ zu ermöglichen (► vgl. Kap. 1.1). Denn bei dem Datentreuhandmodell wird die Verwaltung der personenbezogenen Daten an Treuhänder*in abgegeben. Zwar handeln Treuhänder*innen im Sinne der Treugebenden aber wie oben festgestellt wurde, sind Daten keine statischen Produkte. Sie verändern sich jedes Mal, wenn sie informationstechnisch verarbeitet werden. Bei jeder Verarbeitung der Daten müssten Treuhänder*innen die Zustimmung der Datenbesitzenden einholen. Die Selbstbestimmung und Selbstverwaltung entfallen dann mit dem Datentreuhandmodell. Dies macht deutlich, dass die Beteiligung der Zivilgesellschaft an den strategischen Datenmaßnahmen der Bundesregierung einbezogen und ermöglicht werden sollte.

Die Untersuchung an dieser Stelle ist abgeschlossen. Im Folgenden wird mit dem Fazit die vorliegende Arbeit vollendet.

6. Fazit

Das Ziel der vorliegenden Arbeit bestand darin, eine systematische Literaturanalyse mit qualitativer Inhaltsanalyse durchzuführen, um zum einen die Ursachen von pre-existing Bias in Daten zu untersuchen und zum anderen, ob zur Behebung von pre-existing Bias in Daten mögliche Lösungsansätze gibt. Zudem sollte herausgefunden werden, ob es bereits Konzepte oder Modelle gibt, die pre-existing Bias in Daten beheben. Hierzu wurde die vorliegende Arbeit folgendermaßen aufgebaut.

Zuerst wurden die Relevanz und die Motivation für die vorliegende Arbeit dargelegt. Als Nächstes wurden zentrale Begriffe definiert und der Forschungsstand zum Themenkomplex dargestellt sowie der theoretische Bezugsrahmen festgesetzt. Darauffolgend wurde im Kapitel Methodik die Herleitung der Forschungsfrage aufgezeigt, sowie die systematische Vorgehensweise der Literaturrecherche dargelegt. Zum Abschluss dieses Kapitel wurden die Methoden zur Datenextraktion aufgezeigt sowie die Auswertungsmethoden mittels qualitativer Analyse dargestellt. Im nächsten Kapitel werden die Forschungsfragen beantwortet und im Anschluss daran werden die Forschungsergebnisse zusammenfassend dargestellt. Zudem wird eine Reflexion über die Untersuchung für die vorliegende Arbeit durchgeführt und mit einem Ausblick beendet.

6.1 Beantwortung der Forschungsfragen

Die Beantwortung der Forschungsfragen ist in drei Kapiteln unterteilt. Das erste Kapitel beantwortet die Forschungsfrage: Was sind die Ursachen von pre-existing Bias in Daten? Dabei wird zunächst allgemein auf die Frage eingegangen und im Unterkapitel spezifizieren sich die Antworten auf die Themen Datenqualität, Datensammlung, Datenerhebung und Datengrundlage. Gleiches gilt für die zweite Forschungsfrage: Was ist in der wissenschaftlichen Literatur zur Behebung von pre-existing Bias in Daten bekannt? Im dritten Kapitel werden Konzepte und Modelle vorgestellt, die das Problem pre-existing Bias mit neuen Ansätzen lösen.

6.1.1 Ursachen von pre-existing Bias in Daten

Die SLR ließ erkennen, dass die Ursachen von pre-existing Bias in Daten die DNA einer Gesellschaft in sich tragen. Denn die genetischen Informationen einer Gesellschaft sind in Daten gespeichert. Infolgedessen offenbaren sich Diskriminierungspraktiken einer Gesellschaft in der algorithmischen Informationsverarbeitung einer KI-Anwendung. So kann strukturelle Diskriminierung in KI-Anwendungen hervorgerufen und verstärkt werden, wenn soziale Gruppen in Kategorien aufgeteilt werden, die unterschiedliche Statusgewichtungen erhalten bzw.

über- oder unterrepräsentiert sind. Was zur Folge hat, das soziale Gruppen, die einer größeren Population angehören im Entscheidungsprozess einer KI-Anwendung auch besser bewertet werden. Denn eine KI-Anwendung erlernt das Wissen über die ihr zur Verfügung gestellten Daten. Im Kern entspricht das Handeln von KI-Anwendungen den Prinzipien des Utilitarismus. Das Konzept des Utilitarismus ist auf das Allgemeinwohl einer Gesellschaft sowie auf dessen Maximierung ausgerichtet. Auch eine KI-Anwendung handelt nach dem utilitaristischen Prinzip. Eine KI-Anwendung prozessiert Daten durch statistisches Verfahren, welches die Datenverarbeitung über die größtmöglichen Populationen klassifiziert, um allgemeingültige Aussagen zu treffen. Die Darstellung sowie Größe von Populationen in Datensätzen sind daher entscheidend, wenn KI-Anwendungen Gerechtigkeit walten und faire Entscheidungen treffen sollen. Daher werden bereits im Designprozess einer KI-Anwendung das Fundament für eine gemeinwohlorientierte KI-Anwendung gelegt. Denn fehlendes Verständnis über die Datensätze führt zu Verzerrungen oder aber nicht gewolltes Verständnis. Denn Daten sind die Währung zur Machterhaltung und Big-Tech Konzerne sind an Daten interessiert, die sie als Währung zur Machterhaltung einsetzen. Ihre gutgemeinten Absichtserklärungen KI-Anwendungen gerecht zu gestalten, bleiben ein Lippenbekenntnis, das sich nur dann ändern wird, wenn es um die Vorteile der eigenen (männlichen) Machterhaltung geht oder eben diese gefährdet sind.

6.1.1.1 Datenqualität

Die SLR ließ erkennen, dass die Ursachen für die Datenqualität zum einen durch die Methoden der Datenaufbereitung verursacht werden. Zum anderen beeinflusst die Datenqualität die unhinterfragte Weltanschauung bei der Datenverwendung. Denn Daten spiegeln die Aspekte der Realität wider und werden von denjenigen wahrgenommen, die an der Entstehung der Datensätze beteiligt sind. Demnach fließen die Perspektiven und Weltanschauung von einer kleinen homogenen Gruppe von Menschen in die Datenqualität ein und manifestieren sich in KI-Anwendungen. Dadurch entsteht mangelnde Perspektivenvielfalt, die ein kritischer Faktor für die KI-Entwicklung ist. Das Thema Diversität wird zwar als zentrale Transformation für die Gesellschaft betrachtet, jedoch nur, wenn Organisation einen Mehrwert für Innovationen durch diverse Teammitglieder erkennen. Diversität bleibt ein Lippenbekenntnis und Inklusion ist eine Frage der Teilhabe, wenn unterrepräsentierte und gefährdete Bevölkerungsgruppen nicht an der Entwicklung von Datentechnologien beteiligt werden und Mitbestimmungsrechte erhalten, dann werden Unterdrückungsmechanismen sich normalisieren. Ferner ließ die SLR erkennen, dass Diskriminierungen in KI-Anwendungen durch Daten verursacht werden. Im Verarbeitungsprozess der Daten kodieren sich historische Ungerechtigkeiten, deren Voreingenommenheit als nicht zu

vermeidendes Merkmal in Daten hervortreten. Um dieses Dilemma zu lösen, werden Bestrebungen in der Informatik angestellt, unvoreingenommene und diskriminierungsfreie Algorithmen zu entwickeln oder die mangelnden Perspektiven in der KI-Entwicklung zu umgehen, in dem technische Lösungen für diskriminierungsfreie Datensätze gesucht werden. Es wird der Mythosabbau in Bezug auf die Objektivität von Algorithmen gefordert, die Fairness durch technische Lösungen implizieren. Zudem existieren unterschiedliche Definitionen für Fairness und Voreingenommenheit in der KI-Welt. Durch diese Vielfalt an Definitionen ist zum einen kein standarisiertes einheitliches Vorgehen in der Modellierung von Algorithmen erkennbar und zum anderen erweckt es den Eindruck, dass für Gerechtigkeit verschiedene Bedeutungen geben würde, die durch heuristisches Vorgehen verschiedene Fairnessmetriken hervorrufen. So werden Fairness-Toolkits entwickelt, die verschiedene Entzerrungsmethoden beinhalten, mit denen soziale Gerechtigkeit in KI-Anwendungen durch mathematische Operationalisierung erreicht werden soll. Verschiedene Untersuchungen zeigen jedoch auf, dass die Daten von unterschiedlichen sozialen Gruppen wie bspw. die LGBTQ+-Community, Kinder, Hochschulzulassungen von Studienbewerbenden, ältere Erwachsene usw. unterschiedliche Lebenswelten darstellen und bei der Analyse der Fairness von KI-Anwendungen einen angemessenen Granularitätsgrad erfordern. Daher werden Abhilfemaßnahmen vorgeschlagen, die bei Menschen und Prozessen in der KI-Entwicklung sowie Datenaufbereitung angesetzt werden sollten, um die Lebenswelten von verschiedenen sozialen Gruppen zu verstehen.

6.1.1.2 Datensammlung

Die SLR ließ erkennen, dass Datensammlungen und Datenkuration mit Übereinstimmung einer Ideologie oder Weltanschauungen erfolgen. Denn u.U. ist es gar nicht möglich Daten zu sammeln, ohne ein Unterscheidungspraxis anzuwenden und zu dem soll noch der Schutz der Privatsphäre berücksichtigt werden. Die Daten von Menschen werden aufgrund von sichtbaren Merkmalen gesammelt. Selbst Namen oder Adressen geben Rückschlüsse auf geschützte Merkmale. Daher führt jede Datensammlung einen politischen Akt aus, der durch KI-Anwendungen durchgesetzt wird. Ferner ließ die SLR erkennen, dass umfangreiche und sensible Daten gesammelt werden, ohne dass die Bedingungen für eine informierte Zustimmung bekannt sind. Die gesammelten Daten werden ausgewertet und gegen die Anwendenden eingesetzt, um sie bei ihrer Entscheidungsfindung zu manipulieren. Zudem zeigt sich, dass die Datensätze von schutzbedürftigen und unterrepräsentierten Gruppen nach wie vor systembedingten strukturellen Verzerrungen ausgesetzt sind. Es sind praktisch alle Datensätze von Personen betroffen, die nicht in der KI-Community vertreten sind oder deren Mitgliedern ähnlich sind.

Die SLR ließ erkennen, dass die größten Akteur*innen im KI-Bereich gerade einmal neun Unternehmen mit Sitz in China und den USA sind, die die Datafizierung maßgeblich vorantreiben. Denn durch die Datenerzeugung und Auswertung verschaffen sie sich nicht nur finanzielle Vorteile, sondern verbessern durch die Datafizierung die Vorhersagen des ML einer KI-Anwendung. Von den finanziellen Vorteilen wollen die Regierungen auch profitieren und versprechen sich durch die Konzeptualisierung von Daten ökonomisches Wachstum. Der Wachstumsanspruch der westlichen Regierungen steht jedoch im Widerspruch zum Grundrecht des Einzelnen auf Privatsphäre und führt zu einem Paradoxon der Ethik- und Fairnessansprüche. Zudem kritisieren Publizierende, dass die kolonialen Praktiken sich wiederholen, indem die Länder des Globalen Südens die Versorgungskette der KI-Anwendungen sicherstellen.

Zudem ließ die SLR erkennen, dass durch die Datafizierung neue Wege zur Überwachungspraktiken und Informationskontrollen eröffnen. Überwachungspraktiken, die zu Unterdrückungsmechanismen führen, werden vordergründig dadurch argumentiert, dass zur Fairnesseinhaltung von KI-Anwendungen noch mehr Daten von unterrepräsentierten sozialen Gruppen erhoben und gesammelt werden sollen, um die Klassifizierungsergebnisse zu verbessern, wobei für die überrepräsentierten Gruppen keine weitere Datenerhebung und -sammlung stattfindet. Dadurch werden die überrepräsentierten Gruppen als die Norm angenommen, während die unterrepräsentierten sozialen Gruppen die Abweichung von der Norm sind. Mit der Informationskontrolle ist gemeint, dass Big-Tech Konzerne eine Monopolstellung erlangt haben, die den Zugang zu Informationen bestimmen.

6.1.1.3 Datenerhebung

Die SLR ließ erkennen, dass Datenerhebungen die Weltanschauung des Liberalismus in westlichen Ländern ins Wanken bringen. Denn einerseits soll die freie Entfaltung und Autonomie von Personen in einer Gesellschaft sichergestellt werden, andererseits soll Wirtschaftswachstum durch digitale personenbezogene Daten generiert werden. Dieses Paradoxon sollen KI-Anwendungen erfüllen, indem sie die Fairnessansprüche im Einklang mit den Konzepten der Gerechtigkeitstheorie erfüllen. Es wird festgestellt, dass die ethischen Anforderungen mit den politischen und ökonomischen Dimensionen weitestgehend unvereinbar sind. Ferner wird festgestellt, dass eine KI-Anwendung insoweit ethisch sein wird wie der Zweck des sozialen Systems, das sie einsetzt und betreibt. Darüber hinaus ließ die SLR erkennen, dass KI-Anwendungen als soziotechnisches System die Bestätigung derjenigen suchen sollten, die am meisten von ihnen betroffen sind. Mediale Diskurse beinhalten jedoch die Erfolge der Wirtschaft, die mit KI-Anwendungen Gewinne erzielen. Die Zivilgesellschaft und die Politik sind kaum an diesen Diskursen beteiligt. Es wird lediglich von ihnen der

informierter Umgang mit digitalen Artefakten erwartet. Zudem ließ die SLR erkennen, dass Datenmonetarisierung im Hinblick auf die Rechte an geistigem Eigentum von personenbezogenen Daten den liberalen Theorien zuwiderlaufen. Denn geistiges Eigentum ist veräußerbar, die Privatsphäre jedoch nicht. Darüber hinaus ließ die SLR erkennen, dass das Spannungsverhältnis zwischen der Einhaltung des Datenschutzrechts und die Erhebung personenbezogener Daten bestehen bleibt. Zwar fallen algorithmische Entscheidungsprozesse nicht in den Anwendungsbereich des Datenschutzrechts, wenn personenbezogene Daten anonymisiert sind. Jedoch wirkt der Datenschutz nicht, wenn Vorhersagen für eine Gruppe oder Region durch algorithmische Entscheidungsprozesse gemacht werden. Zusätzlich haben sich Erkenntnisse durch die SLR ergeben, dass Datenlabeling in hyperspezialisierte Aufgaben aufgeteilt ist. Diese Mikroaufgaben werden von den sog. Crowdworkern ausgeführt, die über die Bedeutung ihrer Arbeit nicht in Kenntnis gesetzt werden. Bei der Kennzeichnung der Daten werden zudem den Crowdworkern feste Vorgaben gegeben, die die differenzierte Bezeichnungen der Daten nicht zulassen. Die differenzierte Kennzeichnung der Daten ist jedoch notwendig, um Diskrepanzen aufzudecken.

6.1.1.4 Datengrundlage

Die SLR ließ erkennen, dass eine undifferenzierte Betrachtungsweise auf die Datengrundlage die Leistung der KI-Anwendungen beeinträchtigen. Demnach berücksichtigt eine differenzierte Betrachtungsweise der Datengrundlage die vielfältigen Eigenschaften von Menschen wie bspw. Geschlecht, Alter, Gesichtsbehaarung usw. sowie deren Lebenswelt. Zudem ließ die SLR erkennen, dass wirksame methodische Instrumente fehlen, um die Trainingsdaten im Hinblick auf die demografischen Merkmale einer Stichprobengruppe zu bewerten.

Ferner werden Ansätze zum Ausschluss von sensiblen Daten in der Datengrundlage vorgeschlagen, die als Trainingsdaten den KI-Anwendungen zur Verfügung gestellt werden. Es wird festgestellt, dass dieser Ansatz, auch „fairness through unawareness“ genannt, die Unterdrückungsmechanismen der „*weißen* Vorherrschaft“ verdecken und die Opazität der Identifizierungsmechanismen von Verzerrungen erhöhen, wenn die Kategorie „*Race*“ aus den Datensätzen entfernt bzw. zur Klassifizierung nicht berücksichtigt wird. KI-Anwendungen erleiden dann an „color blindness“, die die strukturelle Diskriminierung gegenüber unterrepräsentierten sozialen Gruppen verstärken.

Ferner ließ die SLR erkennen, dass Verzerrungen in KI-Anwendungen eine Ähnlichkeit mit der Erziehung von Kindern haben. Wenn KI-Anwendung in der Lernphase aus den Trainingsdaten strategische Täuschung, Ignoranz, kognitive Dissonanz, Gewalt

und eine Reihe anderer nicht hilfreicher oder destruktiver menschlicher Tendenzen das Wissen sich aneignen, dann ist etwas in der Erziehung falsch gelaufen.

Der folgende Abschnitt beantwortet die zweite Forschungsfrage und thematisiert das Themenkomplex Behebungen von pre-existing Bias.

6.1.2 Behebungen von pre-existing Bias

Die SLR ließ erkennen, dass zur Behebung von pre-existing Bias ein gesellschaftlicher Diskurs notwendig ist. Denn die Freiheit der individuellen Autonomie wird durch gemeinwohlorientierte KI-Anwendungen in den Hintergrund gedrängt, während die Aufmerksamkeit auf die kollektive Autonomie gelenkt wird. Es wird für die Durchsetzung der kollektiven Interessen vorgeschlagen, dass ein Rahmenwerk zur Einbindung der Gesellschaft in die KI-Forschung festlegt werden sollte sowie die Schaffung von Anreizmöglichkeiten für KI-Forschende, die betroffene Gemeinschaften einbeziehen.

6.1.2.1 Datenqualität

Für die Verbesserung von Datenqualität wird Data Auditing vorgeschlagen, die entweder im Rahmen einer algorithmischen Prüfung stattfindet oder für sich allein. Ferner wird für die Verbesserung der Datenqualität die Methode der Datenerweiterung durch synthetische Datenpunkte vorgeschlagen. Hierbei wird ein ‚Idealwelt-Datensatz' konstruiert, der ungleich verteilte Attribute mit synthetischen Datenpunkten ausgleicht.

Zudem ließ die SLR erkennen, dass Diversität in der KI-Entwicklungen zur Gestaltung von fairen KI-Anwendungen beitragen kann. Hierbei werden Verbesserungspotentiale in der Vernetzung von Hochschuleinrichtungen, Einbeziehung von Interessenvertretenden aus verschiedenen Gemeinschaften in den Forschungsprozess und Förderung von diversen Gruppen künftiger Führungskräfte innerhalb der KI-Forschungsaktivitäten gesehen. Ferner ließ die SLR erkennen, dass durch das Einbeziehen von marginalisiertem Wissen die Datenqualität verbessert wird. Denn durch die Inklusionsmechanismen wird die Demokratisierung der KI-Anwendungen gefördert. Außerdem ließ die SLR erkennen, dass Schulungsmaßnahmen für KI-Ingenieur*innen zur Behebung von pre-existing Bias eine wirkungsvolle Maßnahme sind, wenn sie Zugang zur Rechtsberatung erhalten sowie Schulungen zu Soft Skills und ein allgemeines Verständnis über den Einsatz von soziotechnischen KI-Anwendungen als Schulungsmaßnahme erhalten.

Des Weiteren ließ die SLR ließ erkennen, dass die Datenqualität sich verbessern könnte, wenn interdisziplinäre Teams in KI-Entwicklungen etabliert werden, die sich mit der Thematik der Datenlieferkette beschäftigen. Das Einbringen der verschiedenen Fachdisziplinen aus Politik, Informatik und Sozialwissenschaft ist der erste Schritt den

Menschen hinter den mathematischen Konstrukten zu sehen. Insbesondere wird die Zusammenarbeit mit dem Fachbereich Sozialwissenschaften befürwortet, die mit theoriegestützten Ansätzen die Mechanismen der strukturellen Diskriminierung bewerten. Speziell werden für die Datenkuration Kenntnisse über Intersektionalität, Feminismus, Gender und Rassismus gefordert sowie antirassistische Haltung. Auch die Installation von Kontrollinstanzen innerhalb und zwischen den Institutionen wäre eine weitere Maßnahme, um die Datensammelnden zu kontrollieren. Derzeit gibt es derartige Sicherheitsvorkehrungen nicht. Eine Alternative hierzu könnte sein, dass Gleichstellungs- und Datenschutzbehörden über Ermittlungs- und Durchsetzungsbefugnisse verfügen. Überdies ließ die SLR erkennen, dass partizipative Designansätze ein möglicher Ansatz in der KI-Entwicklung sind, um KI-Entwickelnde und Betroffene zusammenzubringen. Dazu werden u.a. die kollaborativen Ansätze von Participatory Design bzw. Co-Design, Design Justice, Design Thinking und Value Sensitive Design vorgeschlagen. Im Designprozess dieser Ansätze sind die Stimmen aller Beteiligten gleichberichtigt und es existieren keine hierarchischen Ebenen.

6.1.2.2 Datensammlung

Für die Unterkategorie Datensammlung wurde durch die qualitative Inhaltsanalyse in den Publikationen keine Hinweise zur Behebung gefunden. Daher entfällt diese Unterkategorie in der SLR.

6.1.2.3 Datenerhebung

Die SLR ließ erkennen, dass für die Modellierung von gemeinwohlorientierten KI-Anwendungen die Notwendigkeit besteht, den Erfassungsprozess und etwaige Sicherheits- oder Datenschutzbedenken von Datensätzen sowie die dazugehörigen Metadaten zu verstehen, um Kategorisierungen von sensiblen Merkmalen durchzuführen. Zudem sollte der Zweck der Datenerhebung geklärt werden, vornehmlich dann, wenn Daten aus der Datensphäre verwendet werden. Daten direkt bei Betroffenen zu erheben wäre ein möglicher Ansatz, der zum einen die Akzeptanz der KI-Anwendungen erhöhen, und zum anderen die Lizenzierungsbedingungen klären könnte. Um der Transparenz und Rechenschaftspflicht nachzukommen, wird die Verankerung der Dokumentationspflicht in der KI-Entwicklungspipeline gefordert sowie die explizite Datenkennzeichnung, um implizite Annahmen vorzubeugen und die Veröffentlichung der erhobenen Daten. Es wird festgestellt, dass Datenerhebungen ständige Reflexion und Überlegung benötigen, weil Daten an sich kein statischer Gegenstand sind und in Folge ihrer Verarbeitung sich Veränderungsprozessen unterziehen. Um die ethischen Richtlinien rund um personenbezogene Daten zu erhöhen, ist ein möglicher Ansatz die Datenerhebung zu professionalisieren. Dabei könnten die Durchsetzungsstrategien und Lizenzbedingungen von Archiv- und Bibliothekswissenschaften als Vorbild für den Beruf

der Datenerhebenden dienen. Weitere Möglichkeiten wären die Auseinandersetzung mit verschiedenen Organisationsformen, um gemeinwohlorientierte KI-Anwendungen zu entwickeln, die als soziales System von den Mitgliedern selbst verwaltet und kontrolliert werden sowie die eigenständige Einnahmengenerierung.

6.1.2.4 Datengrundlage

Die SLR ließ erkennen, dass es drei Verfahrensarten gibt, die zur Verringerung von Verzerrungen in KI-Anwendungen angewendet werden. Zudem werden verschiedene Methoden zur Datenaufbereitung und Dateneinsatz aufgezeigt, die das Verzerrungsrisiko mindern sollen. Darüber hinaus werden Überlegungen angestellt, wie die Fairness-Zertifizierung für personenbezogene Daten mit operationalisierbaren Anforderungen etabliert werden kann. Best Practices sind durch NGO‘s oder Forschende einer Universität entstanden, die ungerechte KI-Anwendungen von Unternehmen mit eigenen Mitteln untersucht haben. Demnach wäre die Überlegung wert, dass solche Einrichtungen die Evaluierung in Form von Audits und Zertifizierung von soziotechnischen KI-Anwendung als Unabhängige nicht gewinnorientierte Stellen durchführen und entsprechende Ressourcen erhalten.

Um die Datengrundlage zu verbessern, ließ die SLR erkennen, dass disaggregierte Evaluierungen von Datensätzen eine Möglichkeit bieten, Trainingsdaten in ML zu untersuchen.

Ferner ließ die SLR erkennen, dass eine reflexive Haltung gegenüber den Datensätzen in KI-Anwendungen die Datengrundlage verbessern kann. Die reflexive Haltung - unterfüttert mit den kritischen Theorien des Feminismus, Rassismus und Gender - erkennt den dynamischen Veränderungsprozess von Daten und die in Daten kodierten gesellschaftlichen Perspektiven auf soziale Gruppen. In einer wünschenswerten Datenutopie sprechen Datensätze in der Sprache Gleichheit, Freiheit und Emanzipation, die feministische KI-Anwendungen hervorbringen.

Zusätzlich ließ die SLR erkennen, dass ein Perspektivenwechsel auf KI-Anwendungen das Narrativ und das Selbstverständnis der KI-Forschung verändern könnte, indem die Theorie der radikalen Elternschaft einbezogen wird. Hierbei wird festgestellt, dass KI-Anwendungen bisher von Elternschaft entwickelt wurden, die männlich geprägt ist. Die Einbeziehung der weiblich geprägten Elternschaft würde die KI-Technologie transformieren. Wobei hier nicht das Geschlecht an sich gemeint ist, sondern die Rollen. Zudem könnte das Wissen über evolutionäre Organisationsformen für die KI-Entwicklung von Bedeutung sein, um die gesellschaftlichen Machtstrukturen zu dekonstruieren.

Ferner ließ die SLR erkennen, dass radikale Erziehungspraktiken das Verständnis über Trainingsdaten verbessern würden, die in KI-Anwendungen eingesetzt werden. Denn Kinder oder KI-Modelle Erlernen ohne Sorgfalt und Anleitung die Schattenseiten der realen Welt und wenden das erlernte Wissen an.

Das nächste Kapitel beantwortet die dritte Forschungsfrage und thematisiert das Themenkomplex Konzepte und Modelle, pre-existing Bias beheben.

6.1.3 Konzepte oder Modelle

Die SLR hat ergeben, dass verschiedene Organisationsmodelle etabliert werden, die als Widerstand gegen die Hegemonie des Silicon Valleys zum Überwachungskapitalismus entwickelt wurden. Deutlich wird bei den Projekten und Modellen, dass der Top-Down Führungsstil mit Anreizen und Sanktionen durch kollaborative, sinnstiftende und kooperative Organisationsmodelle ersetzt werden. Im Vordergrund der Datenmodelle steht die ganzheitliche Berücksichtigung der sozialen Aspekte einer Gemeinschaft und die Einbeziehung sowie Inklusion aller Beteiligten. Mitglieder sind ausgestattet mit Teilhabemöglichkeiten, die den feministischen Grundsätzen entsprechen.

6.2 Zusammenfassung

KI-Anwendungen, die mit ML aus personenbezogenen Daten lernen, spiegeln nicht nur die Ansichten einer Gesellschaft über soziale Gruppen wider, sondern auch ihre historischen und aktuellen Werte, die innerhalb der Gesellschaft vorherrschen. Dabei spielen Daten die Hauptrolle in KI-Anwendungen, die für die Funktionsweise einer KI-Anwendung verantwortlich sind. Die Wissensaneignung einer KI-Anwendung geht über die Trainingsdaten, die ihr zur Verfügung gestellt werden. In diesen Daten sind Ein- und Ausschlussmechanismen einer Gesellschaft kodiert, die zu struktureller Bevorzugung oder Benachteiligung von sozialen Gruppen führen. Um die strukturellen Diskriminierungspraktiken in der Gesellschaft aufzulösen bzw. zu umgehen, werden technische Lösungen gesucht, die Gerechtigkeit über Fairnessmetriken in KI-Modellen regeln wollen. Jedoch stoßen mathematische Konstrukte in Form von Algorithmen an ihre Grenzen, wenn das soziale Gefüge einer Gesellschaft operationalisiert werden soll. Einer der herausstechenden Erkenntnisse dieser Studie ist, dass die anhaltenden Problematiken rund um KI-Entwicklungen auf Perspektivenmangel zurückzuführen sind. Aufgrund der homogenen Beteiligten in der KI-Entwicklung entstehen blinde Flecken, die durch den eigenen privilegierten Status das Bedürfnis anderer übersehen. Diversitäts- und Inklusionsbestrebungen von Unternehmen als auch von Organisationen verbleiben in gutgemeinten Absichtserklärungen. Mögliche Ansätze zur Diversitäts- und Inklusionsbestrebungen könnten verschiedene partizipative Designansätze sein, die die

Möglichkeit bieten, gemeinwohlorientierte KI-Anwendungen zu entwickeln. Deutlich wird, dass der Beteiligungsprozess von sozialen Gruppen an den Technologieentwicklungen den Perspektivenmangel minimieren könnte, um unvoreingenommene und diskriminierungsfreie KI-Anwendungen zu entwickeln. Auch sind interdisziplinäre Teambildungen eine mögliche Maßnahme, um Diskriminierungen durch KI-Anwendungen vorzubeugen. Insbesondere werden Sozialwissenschaften hervorgehoben, die theoriegeleitet strukturelle Diskriminierung bewerten und zur Datenaufbereitung beitragen können. Auch die Archiv- und Bibliothekswissenschaften, die sich mit Lizenzbedingungen und Dokumentenverwaltung auskennen, könnten zur Datenaufbereitung beitragen.

Für gemeinwohlorientierte KI-Anwendungen sind kooperative und kollaborative Organisationsformen geeignet, die selbstbestimmt und selbstverwaltend geführt werden, wohingegen in Datentreuhandmodellen, so wie im Kap. 1.1 beschrieben, die Treuhänder*innen die Verwaltung der personenbezogenen Daten übernehmen und im Sinne der Datenbesitzenden handeln. Es sollte hierbei nicht außer Acht gelassen werden, das Daten in ihrer Eigenschaft dynamisch sind, und sich in Folge der Datenverarbeitung jedes Mal ändern. Zu ihren Eigenschaften gehören auch, dass sie gleichzeitig genutzt werden können und keine Verschleißeffekte haben. Deshalb ist die Auffassung allgemein verbreitet, dass Daten als Wirtschaftsgut veräußerbar seien. Es werden Überlegungen angestellt personenbezogene Daten zu monetarisieren, weil sie wie geistiges Eigentum angesehen werden könnten. Festgestellt wurde, dass geistiges Eigentum veräußerbar ist, aber nicht die Privatsphäre. Außerdem lässt sich feststellen, dass die liberale Weltanschauung der westlichen Länder in Konflikt steht mit den ethischen Grundsätzen für KI-Anwendungen. Denn einerseits soll die Privatsphäre geschützt werden, andererseits sollen mit personenbezogenen Daten ökonomisches Wachstum generiert werden.

Weitere Erkenntnisse sind die fehlenden Standards in der KI-Entwicklung. Es herrschen diverse Definitionsansätze für den Begriff *Fairness* sowie fehlende Dokumentationsstandards für die verwendeten Methoden und Evaluationsmöglichkeiten von Trainingsdaten in KI-Anwendungen.

Neben anderen Erkenntnissen lässt sich abschließend feststellen, dass bei der Datenkuration eine antirassistische Haltung notwendig ist sowie Kenntnisse der kritischen Theorien über Intersektionalität, Feminismus, Gender und Rassismus. Diese Kenntnisse würden im Prozess der Datenaufbereitung dazu führen, dass Datensätze fair und gerecht gestaltet werden. Demnach würde eine feministische KI-Anwendung die Sprachen der Gleichheit, Freiheit und Emanzipation aus den ihr zur Verfügung gestellten Trainingsdaten erlernen, die diese Sprachen in sich tragen.

6.3 Reflexion der Untersuchungen

Herausforderungen bestanden im Screening Prozess der 258 identifizierten Publikationen. Im ersten Schritt wurden die jeweiligen Abstracts gelesen und nach den Ein- und Ausschlusskriterien bewertet. Der Auswahlprozess der englischsprachigen Publikationen erfolgte bei uneindeutigen Abstracts eher konservativ und wurde in den Auswahlprozess eingeschlossen. Insgesamt fehlten im Screening Prozess aber auch in der qualitativen Inhaltsanalyse weitere Sichtweisen, die bei Uneindeutigkeiten geklärt werden könnten. Um die Replizierbarkeit der systematischen Literaturanalyse zu gewährleisten und die Gütekriterien einer wissenschaftlichen Arbeit zu erfüllen, wurden alle Schritte detailliert protokolliert (► vgl. Kap. 4). Mit der Software MAXQDA wurden die ausgewählten 28 Dokumente analysiert. Dabei haben sich insgesamt 321 Fundstellen ergeben, die auf 59 Seiten in einer separaten Datei ausgelagert wurden.

6.4 Ausblick

Eine feministische KI-Anwendung, die durch maschinelles Lernen ihr Wissen anhand von Trainingsdaten sich aneignet, berücksichtigt alle Aspekte der personenbezogenen Daten. Diese sind in Datensätzen angemessen vertreten und werden nicht aufgrund ihrer beobachtbaren Merkmale durch KI-Anwendungen bewertet oder deren Verhalten vorhergesagt. Feministische KI-Anwendungen sind auf das Gemeinwohl ausgerichtet und berücksichtigen alle Lebenswelten. Ihr oberstes Ziel ist Gerechtigkeit ohne jeglichen Anspruch auf Einnahmengenerierung. Falls Einnahmen generiert werden, dann verteilt eine feministische KI-Anwendung die Einnahmen gerecht unter ihren Mitgliedern. Zudem ermöglichen feministische KI-Anwendungen zivilgesellschaftliche Eigenaktivitäten, welche die höchste Partizipationsstufe der Teilhabemöglichkeit ist. Auf diese Partizipationsstufe organisiert die Gemeinschaft sich selbst und bestimmt, wie Daten der Mitglieder eingesetzt werden und behandelt sie gerecht. Der Staat stellt die entsprechenden Ressourcen für die Selbstorganisation der Gemeinschaft zur Verfügung und hat kein Mitbestimmungsrecht. An der feministischen KI-Entwicklung sind alle Menschen beteiligt, deren Daten in KI-Anwendungen verarbeitet werden. Ihre Verarbeitungsprozesse sind für alle verständlich dokumentiert und die Öffentlichkeit hat Zugang zu den Dokumentationen. Die Datenerhebungen für feministische KI-Anwendungen finden kontextbezogen und direkt mit den Datenbesitzenden statt.

So oder so ähnlich könnte die Zukunft einer gerechten KI-Anwendung gestaltet werden. Eine wünschenswerte Zukunftsvision könnte sein, dass eine feministische KI-Anwendung die Ungerechtigkeiten einer Gesellschaft aushebelt, die in personenbezogenen Daten kodiert sind.

Literaturverzeichnis

Monografien

Alpaydın, Ethem (2016): Machine learning. The new AI. Cambridge, MA: MIT Press (The mit press essential knowledge series).

Alpaydın, Ethem (2022): Maschinelles Lernen. 3., aktualisierte und erweiterte Auflage. Berlin, Boston: De Gruyter Oldenbourg (De Gruyter Studium).

Am Schmidt Busch, Hans-Christoph (2012): Über das weltweite soziale Chaos. Ausgewählte Schriften zur Philosophie und Gesellschaftstheorie. Berlin: Akademie Verlag (Schriften zur europäischen Ideengeschichte, Band 6).

Ammon, Sabine; Beck, Birgit; Dössel, Olaf; Hermann, Isabella; Markschies, Christoph Johannes; Molnár-Gábor, Fruzsina et al. (2021): Verantwortungsvoller Einsatz von KI? Mit menschlicher Kompetenz! Berlin: Berlin-Brandenburgische Akademie der Wissenschaften (#VerantwortungKI - Künstliche Intelligenz und gesellschaftliche Folgen, 4/2021).

Arndt, Susan (2020): Die 101 wichtigsten Fragen. Rassismus. Originalausgabe, 4. Auflage. München: Verlag C.H.Beck (C. H. Beck Paperback, 7036).

Baig, Samira (2010): Diversity-Management zur Überwindung von Diskriminierung? In: Ulrike Hormel (Hg.): Diskriminierung. Grundlagen und Forschungsergebnisse. 1. Aufl. Wiesbaden: VS Verl. f. Sozialwissenschaften, S. 347–360.

Bauberger, Stefan (2020): Welche KI? Künstliche Intelligenz demokratisch gestalten. München: Hanser.

Beauclair, Wilfried (1968): Rechnen Mit Maschinen. Eine Bildgeschichte der Rechentechnik. Wiesbaden: Springer Vieweg. in Springer Fachmedien Wiesbaden GmbH.

Berger, Peter L.; Luckmann, Thomas (2013): Die gesellschaftliche Konstruktion der Wirklichkeit. Eine Theorie der Wissenssoziologie. Unter Mitarbeit von Helmuth Plessner und Monika Plessner. 25. Aufl. Frankfurt am Main: Fischer-Taschenbuch-Verl. (Fischer, 6623).

Bortz, Jürgen; Döring, Nicola (2016): Forschungsmethoden und Evaluation. In den Sozial- und Humanwissenschaften. 5. Aufl. Berlin: Springer-Verlag.

Brandstetter, Nicole; Dobler, Ralph-Miklas; Ittstein, Daniel Jan (Hg.) (2021): Mensch und Künstliche Intelligenz. Herausforderungen für Kultur, Wirtschaft und Gesellschaft.

Burel, Simone; Saur, Franziska; Tsehaye, Wintai (2020): Quick Guide Female Leadership. Frauen in Führungspositionen in der Arbeitswelt 4.0. Berlin, Heidelberg: Springer Gabler (Quick Guide).

Crenshaw, Kimberlé (2019): Das Zusammenwirken von Race und Gender ins Zentrum rücken. Eine Schwarze feministische Kritik des Antidiskriminierungsdogmas, der feministischen Theorie und antirassistischer Politik (1989). Übersetzt von Céline Barry. In: Natasha A. Kelly (Hg.): Schwarzer Feminismus. Grundlagentexte. 1. Auflage. Münster: UNRAST.

Davis, Kathy; Evans, Mary (Hg.) (2011): Transatlantic conversations. Feminism as travelling theory. Farnham: Ashgate (The feminist imagination).

Deutscher Dialogmarketing Verband e.V. (Hg.) (2019): Dialogmarketing Perspektiven 2018/2019. Wiesbaden: Springer Fachmedien Wiesbaden.

D'Ignazio, Catherine; Klein, Lauren F. (2020): Data feminism. Cambridge, Massachusetts, London, England: The MIT Press (<Strong> ideas series).

Dimitriou, Minas; Schweiger, Gottfried (Hg.) (2015): Fairness und Fairplay. Wiesbaden: Springer Fachmedien Wiesbaden.

Dimitriou, Minas; Schweiger, Gottfried (2015): Fairness und Fairplay. Eine interdisziplinäre Annährung. In: Minas Dimitriou und Gottfried Schweiger (Hg.): Fairness und Fairplay. Wiesbaden: Springer Fachmedien Wiesbaden, 15-22.

D'Onofrio, Sara; Meier, Andreas (2021): Big Data Analytics. Wiesbaden: Springer Fachmedien Wiesbaden.

Flach, Peter (2012): Machine learning. The art and science of algorithms that make sense of data. 1. publ. Cambridge: Cambridge Univ. Press.

Floridi, Luciano (Hg.) (2021): Ethics, Governance, and Policies in Artificial Intelligence. 1st ed. 2021. Cham: Springer International Publishing; Imprint Springer (Springer eBook Collection, 144).

Floridi, Luciano; Cowls, Josh (2021): A Unified Framework of Five Principles for AI in Society. In: Luciano Floridi (Hg.): Ethics, Governance, and Policies in Artificial Intelligence. 1st ed. 2021. Cham: Springer International Publishing; Imprint Springer (Springer eBook Collection, 144), S. 5–18.

Frochte, Jörg (2018): Maschinelles Lernen. Grundlagen und Algorithmen in Python. München: Hanser (Hanser eLibrary).

Gerhard, Ute (2020): Frauenbewegung und Feminismus. Eine Geschichte seit 1789. 4., aktualisierte und erweiterte Auflage. München: C.H. Beck (C.H. Beck Wissen, 2463).

Görz, Günther; Schneeberger, Josef; Schmid, Ute (2013): Handbuch der Künstlichen Intelligenz. 5., überab. und korrigierte Aufl. München: Oldenbourg.

Hark, Sabine (Hg.) (2007): Dis/Kontinuitäten: feministische Theorie. 2., aktualisierte und erweiterte Auflage. Wiesbaden: VS Verlag für Sozialwissenschaften (Lehrbuch zur sozialwissenschaftlichen Frauen- und Geschlechterforschung, 3).

Henning, Martin (2021): KI-Marketing und Gesellschaft. In: Nicole Brandstetter, Ralph-Miklas Dobler und Daniel Jan Ittstein (Hg.): Mensch und Künstliche Intelligenz. Herausforderungen für Kultur, Wirtschaft und Gesellschaft, S. 103–121.

Hill Collins, Patricia (2000): Black feminist thought. Knowledge, consciousness, and the politics of empowerment. 2. ed., rev. 10th anniversary ed. New York, NY: Routledge (Perspectives on gender).

Höffe, Otfried (2006): 1. Einführung in Rawls' Theorie der Gerechtigkeit. In: Otfried Höffe (Hg.): John Rawls: Eine Theorie der Gerechtigkeit: Akademie Verlag GmbH, S. 3–26.

Holland-Cunz, B. (2018): Was ihr zusteht – Kurze Geschichte des Feminismus. In: Feministische Geographien.

Holland-Cunz, Barbara (2019): Feministische Demokratiekritik: Geschlechterforschung als Theorie der Demokratisierung. In: Beate Kortendiek, Birgit Riegraf und Katja Sabisch (Hg.): Handbuch interdisziplinäre Geschlechterforschung. Wiesbaden: Springer VS (Geschlecht und Gesellschaft, Band 65).

Hooks, Bell (2000): Feminism is for everybody. Passionate politics. Cambridge, Mass.: South End Press.

Hormel, Ulrike (Hg.) (2010): Diskriminierung. Grundlagen und Forschungsergebnisse. 1. Aufl. Wiesbaden: VS Verl. f. Sozialwissenschaften.

Kelly, Natasha A. (Hg.) (2019): Schwarzer Feminismus. Grundlagentexte. 1. Auflage. Münster: UNRAST.

Kelly, Natasha A. (2021): Rassismus. Strukturelle Probleme brauchen strukturelle Lösungen! Originalausgabe. Zürich: Atrium Verlag.

Kemper, Andreas (2016): Klassismus. Eine Bestandsaufnahme. Erfurt: Friedrich-Ebert-Stiftung, Landesbüro Thüringen.

Kortendiek, Beate; Riegraf, Birgit; Sabisch, Katja (Hg.) (2019): Handbuch interdisziplinäre Geschlechterforschung. Springer Fachmedien Wiesbaden. Wiesbaden: Springer VS (Geschlecht und Gesellschaft, Band 65).

Kurzweil, Ray (2013): Menschheit 2.0. Die Singularität naht. 1. Aufl. Berlin: Lola Books.

Laloux, Frédéric (2015): Reinventing organizations. Ein Leitfaden zur Gestaltung sinnstiftender Formen der Zusammenarbeit. München: Verlag Franz Vahlen.

Laloux, Frédéric (2017): Reinventing Organizations visuell. Ein illustrierter Leitfaden sinnstiftender Formen der Zusammenarbeit. Unter Mitarbeit von Etienne Appert. München: Verlag Franz Vahlen.

Lenz, Ilse (2019): Feminismus. Denkweisen, Differenzen, Debatten. In: Handbuch interdisziplinäre Geschlechterforschung; Band 1.

Lutz, Christiane (2016): Mythen und Märchen in der psychodynamischen Therapie von Kindern und Jugendlichen. 1. Auflage. Hg. v. Hans Hopf und Arne Burchartz. Stuttgart: Kohlhammer Verlag.

Lutz, Helma; Vivar, Maria Teresa Herrera; Supik, Linda (2011): Framing Intersectionality: An Introduction. In: Kathy Davis und Mary Evans (Hg.): Transatlantic conversations. Feminism as travelling theory. Farnham: Ashgate (The feminist imagination), S. 1–22.

Mayring, Philipp (2015): Qualitative Inhaltsanalyse. Grundlagen und Techniken.

Neisser, Ulric (1974): Kognitive Psychologie. 1. Aufl. Stuttgart: Klett (Konzepte der Humanwissenschaften).

Otte, Ralf (2019): Künstliche Intelligenz für Dummies.

Raschke, Joachim (2020): Die Erfindung der modernen Demokratie. Innovationen, Irrwege, Konsequenzen. Wiesbaden, Germany: Springer VS.

Richter, Alexander; Gačić, Tamara; Kölmel, Bernhard; Waidelich, Lukas (2019): Künstliche Intelligenz und potenzielle Anwendungsfelder im Marketing. In: Deutscher Dialogmarketing Verband e.V. (Hg.): Dialogmarketing Perspektiven 2018/2019. Wiesbaden: Springer Fachmedien Wiesbaden, S. 31–52.

Rost, Detlef H. (2013): Handbuch Intelligenz. 1. Aufl. Weinheim, Basel: Beltz.

Scherr, Albert (2010): Diskriminierung und soziale Ungleichheiten. Erfordernisse und Perspektiven einer ungleichheitsanalytischen Fundierung von Diskriminierungsforschung und Antidiskriminierungsstrategien. In: Ulrike Hormel (Hg.): Diskriminierung. Grundlagen und Forschungsergebnisse. 1. Aufl. Wiesbaden: VS Verl. f. Sozialwissenschaften, S. 35–60.

Schiedermeier, Gudrun (2021): Diskriminierende Systeme - Rassismus und Frauenfeindlichkeit in KI-Systemen. In: Nicole Brandstetter, Ralph-Miklas Dobler und Daniel Jan Ittstein (Hg.): Mensch und Künstliche Intelligenz. Herausforderungen für Kultur, Wirtschaft und Gesellschaft, S. 11–24.

Schneier, Bruce (2015): Data und Goliath. Die Schlacht um die Kontrolle unserer Welt: wie wir uns gegen Überwachung, Zensur und Datenklau wehren können.

Straßburger, Gaby; Rieger, Judith (Hg.) (2014): Partizipation kompakt. Für Studium, Lehre und Praxis sozialer Berufe. Weinheim, Basel: Beltz Juventa.

Strauß, Gerhard (1989): Brisante Wörter von Agitation bis Zeitgeist. Ein Lexikon zum öffentlichen Sprachgebrauch. Berlin, New York: Walter de Gruyter (Schriften des Instituts für deutsche Sprache, Bd. 2).

Truth, Sojourner (2019): Bin ich etwa keine Frau*? (1851). Übersetzt von Akilah Güç und Luam Belay. In: Natasha A. Kelly (Hg.): Schwarzer Feminismus. Grundlagentexte. 1. Auflage. Münster: UNRAST, S. 17–18.

Tsamados, Andreas; Aggarwal, Nikita; Cowls, Josh; Morley, Jessica; Roberts, Huw; Taddeo, Mariarosaria; Floridi, Luciano (2021): The Ethics of Algorithms: Key Problems and Solutions. In: Luciano Floridi (Hg.): Ethics, Governance, and Policies in Artificial Intelligence. 1st ed. 2021. Cham: Springer International Publishing; Imprint Springer (Springer eBook Collection, 144), 97-124.

Walsh, Toby (2018): It's alive. Wie künstliche Intelligenz unser Leben verändern wird. Hamburg, Ipswich, Massachusetts: Edition Körber; EBSCO Industries.

Wetterich, Cita; Plänitz, Erik (2021): Systematische Literaturanalysen in den Sozialwissenschaften: Verlag Barbara Budrich.

Witten, Ian; Frank, Eibe; Hall, Mark; Pal, Christopher (2017): Data mining. Practical machine learning tools and techniques. Fourth edition. Amsterdam, Boston, Heidelberg, London, New York, Oxford, Paris, San Diego, San Francisco, Singapore, Sydney, Tokyo: Elsevier Morgan Kaufmann (Morgan Kaufmann series in data management systems).

Wittpahl, Volker (Hg.) (2019): Künstliche Intelligenz. Technologie, Anwendung, Gesellschaft. [1. Auflage]. Berlin, Heidelberg: Springer Vieweg (OPEN).

Zweig, Katharina A. (2019): Ein Algorithmus hat kein Taktgefühl. Wo künstliche Intelligenz sich irrt, warum uns das betrifft und was wir dagegen tun können. Originalausgabe. München: Heyne.

Digitale Dokumente

AAAI (2006): Celebrating 50th Anniversary of Artificial Intelligence Computing: AAAI/IAAI-06 Conference July 11, 2006. Hg. v. Association for the Advancement of **A**rtificial Intelligence. Online verfügbar unter https://aaai.org/Pressroom/Releases/release-06-0711.php, zuletzt geprüft am 19.03.2022.

Adesso (2018): Künstliche Intelligenz – Anwendungen und Fazit. Mehr als Chatbots und neuronale Netze. Online verfügbar unter https://www.adesso.de/de/news/aditorial/aditorial-ausgabe-1-2018/1-kuenstliche-intelligenz-teil-2.jsp, zuletzt geprüft am 02.05.2022.

Ahuja, Piyush (2013): Man and Machine: Questions of Risk, Trust and Accountability in Today's AI Technology. Online verfügbar unter https://arxiv.org/pdf/1307.7127.pdf, zuletzt geprüft am 17.03.2022.

Alghushairy, Omar; Ma, Xiaogang (2022): Data Storage. In: Laurie A. Schintler und Connie L. McNeely (Hg.): Encyclopedia of Big Data. Cham: Springer International Publishing, S. 338–341. Online verfügbar unter https://10.1007/978-3-319-32010-6_323, zuletzt geprüft am 27.06.2022.

Angwin, Julia; Larson, Jeff; Mattu, Surya; Kirchner, Lauren (2016): Machine Bias. There's software used across the country to predict future criminals. And it's biased against blacks. Hg. v. ProPublica. Online verfügbar unter https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing, zuletzt geprüft am 27.04.2022.

Arndt, Susan (2012): Antrittsvorlesung Prof. Dr. Susan Arndt, Universität Bayreuth. "LiteraturWelten. Transkulturelle Anglistik und der 'Racial Turn' ", Antrittsvorlesung von Prof. Dr. Susan Arndt (Englische Literaturwissenschaft und Anglophone Literaturen) an der Universität Bayreuth vom 24. Oktober 2012. Online verfügbar unter https://vimeo.com/66145276, zuletzt geprüft am 29.05.2022.

Beck, Susanne; Grunwald, Armin; Jacob, Kai; Matzner, Tobias (2019): Künstliche Intelligenz und Diskriminierung. Herausforderungen und Lösungsansätze. Hg. v. Lernende Systeme – Die Plattform für Künstliche Intelligenz. Online verfügbar unter https://www.plattform-lernende-systeme.de/files/Downloads/Publikationen/AG3_Whitepaper_250619.pdf, zuletzt geprüft am 21.05.2022.

BmBF (2022): Künstliche Intelligenz. FORSCHUNG. Bundesministerium für Bildung und Forschung. Online verfügbar unter https://www.bmbf.de/bmbf/de/forschung/digitale-wirtschaft-und-gesellschaft/kuenstliche-intelligenz/kuenstliche-intelligenz_node.html#:~:text=F%C3%BCr%20die%20Umsetzung%20der%20KI,f%C3%BCr%20die%20KI%20%2DF%C3%B6rderung%20bereitgestellt., zuletzt geprüft am 20.03.2022.

Bryson, Joanna J. (2019): The Past Decade and Future of AI's Impact on Society. In: *Towards a New Enlightenment? A Transcendent Decade* (Vol. 11), S. 1–35. Online verfügbar unter https://www.bbvaopenmind.com/wp-content/uploads/2019/02/BBVA-OpenMind-Joanna-J-Bryson-The-Past-Decade-and-Future-of-AI-Impact-on-Society.pdf, zuletzt geprüft am 10.05.2022.

BSI (2021): AI Cloud Service Compliance Criteria Catalogue (AIC4). Hg. v. Federal Office for Information Security. Bundesamt für Sicherheit in der Informationstechnik. Online verfügbar unter https://www.bsi.bund.de/SharedDocs/Downloads/EN/BSI/CloudComputing/AIC4/AI-Cloud-Service-Compliance-Criteria-Catalogue_AIC4.pdf;jsessionid=2FD12F7AA6ECD3C9D8214B651872D2A5.internet462?__blob=publicationFile&v=4, zuletzt geprüft am 30.05.2022.

BSI (2021): Kriterienkatalog für KI-Cloud-Dienste – AIC4. Hg. v. Bundesamt für Sicherheit in der Informationstechnik. Online verfügbar unter https://www.bsi.bund.de/DE/Themen/Unternehmen-und-Organisationen/Informationen-und-Empfehlungen/Kuenstliche-Intelligenz/AIC4/aic4_node.html;jsessionid=8DBF1E745AA468ED6D4C9DCF3FA55DE3.internet461, zuletzt geprüft am 02.03.2022.

BTC-ECHO (2022): Decentralized Autonomous Organization (DAO). Online verfügbar unter https://www.btc-echo.de/academy/bibliothek/decentralized-autonomous-organization-dao/, zuletzt geprüft am 10.07.2022.

Buolamwini, Joy (2019): United States House Committee on Oversight and Government Reform. Hearing on Facial Recognition Technology (Part 1): Its Impact on our Civil Rights and Liberties. Written Testimony of Joy Buolamwini Founder, Algorithmic Justice League. Online verfügbar unter https://www.congress.gov/116/meeting/house/109521/witnesses/HHRG-116-GO00-Wstate-BuolamwiniJ-20190522.pdf, zuletzt geprüft am 26.05.2022.

Burkett, Elinor; Brunell, Laura (2021): "feminism". Hg. v. Encyclopedia Britannica. Online verfügbar unter https://www.britannica.com/topic/feminism. Accessed 26 May 2022., zuletzt geprüft am 26.05.2022.

Buss, Sebastian; Nöldeke, Geeske; Becker, Dennis; Blumtritt, Christoph; Daniels, Marcos; Striapunina, Ksenia (2019): DIGITAL ECONOMY COMPASS. 2019. Hg. v. Statista.

Online verfügbar unter https://de-statista-com.ezp.hs-duesseldorf.de/statistik/studie/id/54238/dokument/digital-economy-compass/, zuletzt geprüft am 27.03.2022.

BVDW (2022): Grundlagen der digitalen Ethik. ETHIK-BIBLIOTHEK. Datenbank. Hg. v. Bundesverband Digitale Wirtschaft e.V. Online verfügbar unter https://www.bvdw.org/themen/digitale-ethik/ethik-bibliothek-1/, zuletzt geprüft am 02.03.2022.

BVerfG (1983): Urteil des Ersten Senats vom 15. Dezember 1983 - 1 BvR 209/83, 1 BvR 484/83, 1 BvR 440/83, 1 BvR 420/83, 1 BvR 362/83, 1 BvR 269/83. Bundesverfassungsgericht. Online verfügbar unter https://www.bundesverfassungsgericht.de/SharedDocs/Downloads/DE/1983/12/rs19831215_1bvr020983.pdf?__blob=publicationFile&v=1, zuletzt geprüft am 27.04.2022.

Çaliskan, Aylin; Bryson, Joanna J.; Narayanan, Arvind (2017): Semantics derived automatically from language corpora contain human-like biases. Supplementary Materials for. In: *Science (New York, N.Y.)* 356 (6334). Online verfügbar unter https://10.1126/science.aal4230, zuletzt geprüft am 06.05.2022.

Capgemini (2021): THE KEY TO DESIGNING INCLUSIVE TECH: Creating diverse and inclusive tech teams. Online verfügbar unter https://www.capgemini.com/de-de/wp-content/uploads/sites/5/2021/07/Final-Report-Web-Version-Inclusive-Tech.pdf, zuletzt geprüft am 19.05.2022.

Charta der Vielfalt e.V. (o.J.): Die Urkunde Charta der Vielfalt im Wortlaut. Diversity als Chance - Die Charta der Vielfalt für Diversity in der Arbeitswelt. Online verfügbar unter https://www.charta-der-vielfalt.de/ueber-uns/ueber-die-initiative/urkunde-charta-der-vielfalt-im-wortlaut/, zuletzt geprüft am 04.07.2022.

Cresci, Stefano (2020): A decade of social bot detection. In: *Commun. ACM* 63 (10), S. 1–16. Online verfügbar unter https://10.1145/3409116, zuletzt geprüft am 21.05.2022.

Datenethikkommission (2019): Gutachten der Datenethikkommission. der Bundesregierung. Hg. v. Datenethikkommission der Bundesregierung. Bundesministerium des Innern, für Bau und Heimat. Online verfügbar unter https://www.bmi.bund.de/SharedDocs/downloads/DE/publikationen/themen/it-digitalpolitik/gutachten-datenethikkommission.pdf?__blob=publicationFile&v=6, zuletzt geprüft am 24.05.2022.

David Dao; Hugo W.; Alf-André Walla; VocalFan; Eric Rubiel; Fabian Schreiber et al. (2022): daviddao/awful-ai: Awful AI - 2021 Edition: Zenodo. Online verfügbar unter https://github.com/daviddao/awful-ai, zuletzt geprüft am 14.07.2022.

DFKI (o.J.): German Research Center for Artificial Intelligence. Online verfügbar unter https://www.dfki.de/en/web/about-us/fellows/research-fellows, zuletzt geprüft am 04.07.2022.

Die Bundesregierung: KABINETT BESCHLIESST FORTSCHREIBUNG DER KI-STRATEGIE DER BUNDESREGIERUN. Definition: Begriffsbestimmung Künstliche Intelligenz. KI, Nationale Strategie für Künstliche Intelligenz, AI Made in Germany. Online verfügbar unter https://www.ki-strategie-deutschland.de/home.html, zuletzt geprüft am 29.04.2022.

Die Bundesregierung (2016): Merkel: Wir müssen uns sputen. Video-Podcast. Online verfügbar unter https://www.bundesregierung.de/breg-de/aktuelles/merkel-wir-muessen-uns-sputen-746750, zuletzt geprüft am 18.02.2022.

Die Bundesregierung (2020): Strategie Künstliche Intelligenz der Bundesregierung. Fortschreibung 2020. Hg. v. KI-Made in Europe. Online verfügbar unter https://www.bmwi.de/Redaktion/DE/Publikationen/Technologie/strategie-kuenstliche-intelligenz-fortschreibung-2020.pdf?__blob=publicationFile&v=12, zuletzt geprüft am 04.02.2022.

Die Bundesregierung (2021): Datenstrategie der Bundesregierung. Eine Innovationsstrategie für gesellschaftlichen Fortschritt und nachhaltiges Wachstum. Kabinettfassung, 27. Januar 2021. Online verfügbar unter https://www.bundesregierung.de/resource/blob/992814/1845634/f073096a398e59573c7526feaadd43c4/datenstrategie-der-bundesregierung-downloadbpa-data.pdf?download=1, zuletzt geprüft am 25.05.2022.

Dreyfus, Hubert L. (1972): What Computers Still Can't Do. A Critique of Artificial Reason. Hg. v. The MIT Press Cambridge, Massachusetts, London England. Online verfügbar unter https://terrorgum.com/tfox/books/whatcomputersstillcantdo_acritiqueofartificialreason.pdf, zuletzt geprüft am 07.04.2022.

Drinhausen, Katja; Brussee, Vincent (2021): CHINA'S SOCIAL CREDIT SYSTEM IN 2021. From fragmentation towards integration. Mercator Institute for China Studies

(MERICS CHINA MONITOR). Online verfügbar unter https://merics.org/sites/default/files/2021-06/MERICS%20ChinaMonitor%2067%20Social%20Credit%20System%20final3.pdf, zuletzt geprüft am 26.04.2022.

DUDEN (2022): Fe-mi-nis-mus, der. Online verfügbar unter https://www.duden.de/rechtschreibung/Feminismus, zuletzt geprüft am 20.02.2022.

Durkin, J. (1996): Expert systems: a view of the field. In: *IEEE Expert* 11 (2), S. 56–63. Online verfügbar unter https:// 10.1109/64.491282, zuletzt geprüft am 01.06.2022.

DWDS (2022): „Fairness“, bereitgestellt durch das Digitale Wörterbuch der deutschen Sprache. Online verfügbar unter https://www.dwds.de/wb/Fairness, zuletzt geprüft am 02.06.2022.

Dzodan, Falvia (2011): MY FEMINISM WILL BE INTERSECTIONAL OR IT WILL BE BULLSHIT! 10.10.2011. Online verfügbar unter https://theresearchpapers.org/my-feminism-will-be-intersectional-or-it-will-be-bullshit/, zuletzt geprüft am 26.05.2022.

EU (2016): zum Schutz natürlicher Personen bei der Verarbeitung personenbezogener Daten, zum freien Datenverkehr und zur Aufhebung der Richtlinie 95/46/EG (Datenschutz-Grundverordnung). DSGVO. Online verfügbar unter https://: https://eur-lex.europa.eu/legal-content/DE/TXT/PDF/?uri=CELEX:32016R0679, zuletzt geprüft am 01.05.2022.

eurostat (2022): Employed ICT specialists by sex. Eurostat defines ICT specialists as "workers who have the ability to develop, operate and maintain ICT systems, and for whom ICT constitute the main part of their job". Online verfügbar unter https://appsso.eurostat.ec.europa.eu/nui/show.do?dataset=isoc_sks_itsps&, zuletzt geprüft am 26.05.2022.

Ferrara, Emilio; Varol, Onur; Davis, Clayton; Menczer, Filippo; Flammini, Alessandro (2016): The rise of social bots. In: *Commun. ACM* 59 (7), S. 96–104. Online verfügbar unter https:// 10.1145/2818717, zuletzt geprüft am 09.05.2022.

Fischer, Sarah; Puschmann, Cornelius (2021): Wie Deutschland über Algorithmen schreibt. Hg. v. Bertelsmann Stiftung. Online verfügbar unter https://www.bertelsmann-stiftung.de/fileadmin/files/user_upload/Diskursanalyse_2021_Algorithmen.pdf, zuletzt geprüft am 08.05.2022.

Forni, Amy Ann; van der Meulen, Rob (2017): Gartner Identifies Three Megatrends That Will Drive Digital Business into the Next Decade. Online verfügbar unter https://www.gartner.com/en/newsroom/press-releases/2017-08-15-gartner-identifies-three-megatrends-that-will-drive-digital-business-into-the-next-decade, zuletzt geprüft am 20.03.2022.

FRA (2017): Second European Union minorities and discrimination survey. Main results. Luxembourg: Publications Offices of the European Union (EU-Midis II). Online verfügbar unter http://publications.europa.eu/de/publication-detail/-/publication/8320df02-da39-11e7-a506-01aa75ed71a1, zuletzt geprüft am 25.06.2022.

Friedler, Sorelle A.; Scheidegger, Carlos; Venkatasubramanian, Suresh; Choudhary, Sonam; Hamilton, Evan P.; Roth, Derek (2019): A comparative study of fairness-enhancing interventions in machine learning. Online verfügbar unter https://arxiv.org/pdf/1802.04422.pdf?ref=https://githubhelp.com, zuletzt geprüft am 24.05.2022.

Friedman, Batya; Nissenbaum, Helen (1996): Bias in Computer Systems. Online verfügbar unter https://dl.acm.org/doi/pdf/10.1145/230538.230561, zuletzt geprüft am 22.05.2022.

Fritsch, Katrin; von Schwichow, Helene (2021): Künstliche Intelligenz und Feminismus: Aktuelle Debatten. Eine Expertise im Rahmen des Projekts KI Thinktank Female Entrepreneurship (KITE). MOTIF Institute for Digital Culture. Online verfügbar unter https://www.kite-bga.de/fileadmin/documents/MOTIF__KI___Feminismus.pdf, zuletzt geprüft am 05.05.2022.

HEG-KI (2018): Eine Definition der KI: Wichtigste Fähigkeiten und Wissenschaftsgebiete. Hochrangige Expertengruppe für künstliche Intelligenz. Europäische Kommission. Online verfügbar unter https://elektro.at/wp-content/uploads/2019/10/EU_Definition-KI.pdf, zuletzt geprüft am 10.05.2022.

HEG-KI (2019): Ethik-Leitlinien für eine vertrauenswürdige KI. Brüssel: Europäische Kommission. Online verfügbar unter https://op.europa.eu/de/publication-detail/-/publication/d3988569-0434-11ea-8c1f-01aa75ed71a1/language-de/format-PDF, zuletzt geprüft am 24.05.2022.

Hill, Robin K. (2016): What an Algorithm Is. In: Philos. Technol. 29 (1), S. 35–59. Online verfügbar unter https://10.1007/s13347-014-0184-5, zuletzt geprüft am 05.06.2022.

Hoffmann, Anna Lauren (2021): Terms of inclusion: Data, discourse, violence. In: *New Media & Society* 23 (12), S. 3539–3556. Online verfügbar unter https://10.1177/1461444820958725, zuletzt geprüft am 28.03.2022.

Hultquist, Carolynne (2022): Data Fusion. In: Laurie A. Schintler und Connie L. McNeely (Hg.): Encyclopedia of Big Data. Cham: Springer International Publishing, S. 285–286. Online verfügbar unter https://10.1007/978-3-319-32010-6_305, zuletzt geprüft am 15.04.2022.

IBM (2021): IBM Unveils World's First 2 Nanometer Chip Technology, Opening a New Frontier for Semiconductors. New chip milestone to propel major leaps forward in performance and energy efficiency. Online verfügbar unter https://newsroom.ibm.com/2021-05-06-IBM-Unveils-Worlds-First-2-Nanometer-Chip-Technology,-Opening-a-New-Frontier-for-Semiconductors, zuletzt geprüft am 28.03.2022.

IBM Research Trusted AI (2018): AI Fairness 360. This extensible open-source toolkit can help you examine, report, and mitigate discrimination and bias in machine learning models throughout the AI application lifecycle. We invite you to use and improve it. Online verfügbar unter https://aif360.mybluemix.net/, zuletzt geprüft am 24.05.2022.

IMD (2021): World Digital Competitiveness Ranking 2021. Measuring the capacity and readiness of economies to adopt and explore digital technologies for economic and social transformation. Hg. v. World Competitiveness Center. Online verfügbar unter https://www.imd.org/centers/world-competitiveness-center/rankings/world-digital-competitiveness/, zuletzt geprüft am 02.03.2022.

Kadadi, Anirudh; Agrawal, Rajeev (2022): Data Integration. In: Laurie A. Schintler und Connie L. McNeely (Hg.): Encyclopedia of Big Data. Cham: Springer International Publishing, S. 290–294. Online verfügbar unter https://10.1007/978-3-319-32010-6, zuletzt geprüft am 27.06.2022.

Koenecke, Allison; Nam, Andrew; Lake, Emily; Nudell, Joe; Quartey, Minnie; Mengesha, Zion et al. (2020): Racial disparities in automated speech recognition. In: *Proceedings of the National Academy of Sciences of the United States of America* 117 (14), S. 7684–7689. Online verfügbar unter https://10.1073/pnas.1915768117, zuletzt geprüft am 16.05.2022.

Krafft, Tobias D.; Zweig, Katharina A. (2019): TRANSPARENZ UND NACHVOLLZIEHBARKEIT ALGORITHMENBASIERTER ENTSCHEIDUNGSPROZESSE. Ein Regulierungsvorschlag aus sozioinformatischer Perspektive. Hg. v. Verbraucherzentrale Bundesverband e.V. Bundesverband der Verbraucherzentralen und Verbraucherverbände. Online verfügbar unter https://www.vzbv.de/sites/default/files/downloads/2019/05/02/19-01-22_zweig_krafft_transparenz_adm-neu.pdf, zuletzt geprüft am 11.05.2022.

Lenz, Ilse (2018): Was ist Feminismus? Heinrich-Böll-Stiftung e.V. Online verfügbar unter https://www.gwi-boell.de/de/2018/05/25/was-ist-feminismus, zuletzt geprüft am 01.12.2021.

Lenz, Ilse (2018): Was ist Feminismus? Heinrich-Böll-Stiftung e.V. Online verfügbar unter https://www.gwi-boell.de/de/2018/05/25/was-ist-feminismus, zuletzt geprüft am 01.12.2021.

Lighthill, James (1972): Artificial Intelligence: A General Survey. Lighthill Report. Online verfügbar unter http://www.chilton-computing.org.uk/inf/literature/reports/lighthill_report/p001.htm, zuletzt geprüft am 17.03.2022.

Lischka, Konrad; Klingel, Anita; Bertelsmann Stiftung (2017): Wenn Maschinen Menschen bewerten. Hg. v. Bertelsmann Stiftung. Online verfügbar unter https://www.bertelsmann-stiftung.de/fileadmin/files/BSt/Publikationen/GrauePublikationen/ADM_Fallstudien.pdf, zuletzt geprüft am 27.04.2022.

Ma, Xiaogang (2017): Metadata. In: Laurie A. Schintler und Connie L. McNeely (Hg.): Encyclopedia of Big Data. Cham: Springer International Publishing, S. 1–5. Online verfügbar unter https://10.1007/978-3-319-32001-4, zuletzt geprüft am 27.06.2022.

Mahesh, Batta (2020): Machine Learning Algorithms - A Review. In: *International Journal of Science and Research (IJSR)* (Volume 9), Artikel Issue 1, S. 381–386. Online verfügbar unter https://www.researchgate.net/profile/Batta-Mahesh/publication/344717762_Machine_Learning_Algorithms_-A_Review/links/5f8b2365299bf1b53e2d243a/Machine-Learning-Algorithms-A-Review.pdf, zuletzt geprüft am 17.04.2022.

Matzner, Tobias (2022): Algorithmen bei Bewerbungsverfahren - sinnvoll und vertretbar? Auf die Frage, wie man einer Ungleichbehandlung von Menschen durch Algorithmen vorbeugen kann und ob Algorithmen, wenn sie richtig designt sind, vorurteilsfreier

urteilen können als Menschen: research in context. Hg. v. science media center germany. Online verfügbar unter https://www.sciencemediacenter.de/alle-angebote/research-in-context/details/news/algorithmen-bei-bewerbungsverfahren-sinnvoll-und-vertretbar/, zuletzt geprüft am 22.05.2022.

McCarthy, J.; Minsky, M. L.; Rochester, N.; Shannon, C.E. (1955): A PROPOSAL FOR THE DARTMOUTH SUMMER RESEARCH PROJECT ON ARTIFICIAL INTELLIGENCE. Online verfügbar unter http://jmc.stanford.edu/articles/dartmouth/dartmouth.pdf, zuletzt geprüft am 15.03.2022.

McNeely, Connie L.; Schintler, Laurie A. (2022): Big Data Concept. In: Laurie A. Schintler und Connie L. McNeely (Hg.): Encyclopedia of Big Data. Cham: Springer International Publishing, S. 79–82. Online verfügbar unter https://10.1007/978-3-319-32010-6_551, zuletzt geprüft am 01.06.2022.

Mehrabi, Ninareh; Morstatter, Fred; Saxena, Nripsuta; Lerman, Kristina; Galstyan, Aram; (2021): A Survey on Bias and Fairness in Machine Learning. In: *ACM Comput. Surv.* 54 (6), S. 1–35. Online verfügbar unter https://0.1145/3457607, zuletzt geprüft am 30.05.2022.

Mittelstadt, Brent Daniel; Allo, Patrick; Taddeo, Mariarosaria; Wachter, Sandra; Floridi, Luciano (2016): The ethics of algorithms: Mapping the debate. In: *Big Data & Society* 3 (2), 205395171667967. Online verfügbar unter https://10.1177/2053951716679679, zuletzt geprüft am 28.03.2022.

Moore, Gordon E. (1965): Cramming more components onto integrated circuits. With unit cost falling as the number of components per circuit rises, by 1975 economics may dictate squeezing as many as 65,000 components on a single silicon chip. In: *Electronics* 1965, 19.04.1965 (Volume 38, Number 8). Online verfügbar unter https://newsroom.intel.com/wp-content/uploads/sites/11/2018/05/moores-law-electronics.pdf, zuletzt geprüft am 28.03.2022.

Neumann, Jana (2018): Leitfaden für die Erstellung einer Abschlussarbeit als systematische Übersichtsarbeit. Universität Duisburg-Essen. Online verfügbar unter https://www.uni-due.de/imperia/md/content/biwi/aopsy/systematische_uebersichtsarbeit-ao.pdf, zuletzt geprüft am 08.05.2022.

Nier, Hedda (2018): Wie weiblich ist die IT? Statista. Online verfügbar unter https://de-statista-com.ezp.hs-duesseldorf.de/infografik/13283/frauen-in-der-tech-branche/, zuletzt geprüft am 01.06.2022.

Olteanu, Alexandra; Castillo, Carlos; Diaz, Fernando; Kıcıman, Emre (2019): Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries. In: *Frontiers in big data* 2. Online verfügbar unter https://10.3389/fdata.2019.0001, zuletzt geprüft am 25.05.2022.

Page, Matthew J.; McKenzie, Joanne E.; Bossuyt, Patrick M.; Boutron, Isabelle; Hoffmann, Tammy C.; Mulrow, Cynthia D. et al. (2021): The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. In: *BMJ (Clinical research ed.)* 372. Online verfügbar unter https://n71. DOI: 10.1136/bmj.n71, zulezt geprüft am 25.05.2022.

Patel, Faiza; Levinson-Waldman, Rachel; DenUyl, Sophia; Koreh, Raya (2019): Social Media Monitoring. SUMMARY: Personal information gleaned from social media posts has been used to target dissent and subject religious and ethnic minorities to enhanced vetting and surveillance. Hg. v. Brennan Center for Justice at New York University School of Law. Online verfügbar unter https://www.brennancenter.org/media/212/download, zuletzt geprüft am 25.05.2022.

Perrault, Ray; Shoham, Yoav; Brynjolfsson, Erik; Clark, Jack; Etchemendy, John; Grosz, Barbara et al. (2019): The AI Index 2019 Annual Report. AI Index Steering Committee, Human-Centered AI Institute, Stanford University. Online verfügbar unter https://hai.stanford.edu/sites/default/files/ai_index_2019_report.pdf, zuletzt geprüft am 21.03.2022.

Perry, Walt; McInnis, Brian; Price, Carter; Smith, Susan; Hollywood, John; (2013): Predictive Policing. The Role of Crime Forecasting in Law Enforcement Operations. Online verfügbar unter https://www.jstor.org/stable/10.7249/j.ctt4cgdcz.9, zuletzt geprüft am 28.04.2022.

Prabhu, Anirudh (2022): Data Discovery. In: Laurie A. Schintler und Connie L. McNeely (Hg.): Encyclopedia of Big Data. Cham: Springer International Publishing, S. 279–283. Online verfügbar unter httpa://10.1007/978-3-319-32010-6, zuletzt geprüft am 29.04.2022.

Reinsel, David; Gantz, John; Rydning, John (2018): The Digitization of the World. From Edge to Core. An IDC White Paper – #US44413318. Online verfügbar unter

https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf, zuletzt geprüft am 17.05.2022.

Russakovsky, Olga; Deng, Jia; Su, Hao; Krause, Jonathan; Satheesh, Sanjeev; Ma, Sean et al. (2015): ImageNet Large Scale Visual Recognition Challenge. IJCV. Online verfügbar unter https://www.image-net.org/about.php; https://www.image-net.org/static_files/files/imagenet_ilsvrc2017_v1.0.pdf, zuletzt geprüft am 19.04.2022.

Saygın, Ayşe Pınar; Çiçekli, İlyas; Akman, Varol (2000): Turing Test: 50 Years Later. In: *Minds and Machines* 10 (4), S. 463–518. Online verfügbar unter https://10.1023/A:1011288000451, zuletzt geprüft am 17.05.2022.

Schintler, Laurie A.; McNeely, Connie L. (Hg.) (2017): Encyclopedia of Big Data. Cham: Springer International Publishing. Online verfügbar unter https://10.1007/978-3-319-32001-4, zuletzt geprüft am 27.06.2022.

Schintler, Laurie A.; McNeely, Connie L. (Hg.) (2022): Encyclopedia of Big Data. Cham: Springer International Publishing. Online verfügbar unter httpa://10.1007/978-3-319-32010-6, zuletzt geprüft am 27.06.2022.

Troncoso, Stacco; Utratel, Ann Marie; McKeon, Timothy; Oñate, Susa; Bryant, Elsie; Bollier, David (2022): If I Only had a Heart: a DisCO manifesto. A joint publication by DisCO.coop, the Transnational Institute and Guerrilla Media Collective. Online verfügbar unter https://disco.coop/wp-content/uploads/2019/11/DisCO_Manifesto-v.1.pdf, zuletzt geprüft am 08.07.2022.

Turing, Alain M. (1950): COMPUTING MACHINERY AND INTELLIGENCE. Online verfügbar unter https://www.csee.umbc.edu/courses/471/papers/turing.pdf, zuletzt geprüft am 14.03.2022.

Veit, Susanne; Koopmans, Ruud; Yemane, Ruta (2018): Ethnische Hierarchien in der Bewerberauswahl: Ein Feldexperiment zu den Ursachen von Arbeitsmarktdiskriminierung. Online verfügbar unter https://www.researchgate.net/publication/327844003_Ethnische_Hierarchien_in_der_Bewerberauswahl_Ein_Feldexperiment_zu_den_Ursachen_von_Arbeitsmarktdiskriminierung, zuletzt geprüft am 22.05.2022.

Vinge, Vemor (1993): THE COMING TECHNOLOGICAL SINGULARITY: HOW TO SURVIVE IN THE POST-HUMAN ERA. NASA. Lewis Research Center, Vision 21:

Interdisciplinary Science and Engineering in the Era of Cyberspace. Online verfügbar unter https://ntrs.nasa.gov/api/citations/19940022856/downloads/19940022856.pdf, zuletzt geprüft am 28.03.2022.

Wissenschaftliche Dienste (2021): Sachstand. Künstliche Intelligenz in der Justiz Internationaler Überblick. Aktenzeichen: WD 7 - 3000 - 017/21; Fachbereich: WD 7: Zivil-, Straf- und Verfahrensrecht, Bau und Stadtentwicklung. Deutscher Bundestag. Online verfügbar unter https://www.bundestag.de/resource/blob/832204/6813d064fab52e9b6d54cbbf5319cea3/WD-7-017-21-pdf-data.pdf, zuletzt geprüft am 27.04.2022.

World Economic Forum (2018): The Global Gender Gap Report. Insight Report. Online verfügbar unter https://www3.weforum.org/docs/WEF_GGGR_2018.pdf, zuletzt geprüft am 01.12.2021.

Zhang, Daniel; Maslej, Nestor; Brynjolfsson, Erik; Etchemendy, John; Lyons, Terah; Manyika, James et al. (2022): The AI Index 2022 Annual Report. AI Index Steering Committee, Stanford Institute for Human-Centered AI, Stanford University. Online verfügbar unter https://aiindex.stanford.edu/wp-content/uploads/2022/03/2022-AI-Index-Report_Master.pdf, zuletzt geprüft am 21.03.2022.

Zweig, Katharina A. (2018): Wo Maschinen irren können. Impuls Algorithmenethik #4. Hg. v. Bertelsmann Stiftung. Online verfügbar unter https://www.bertelsmann-stiftung.de/fileadmin/files/BSt/Publikationen/GrauePublikationen/WoMaschinenIrrenKoennen.pdf, zuletzt geprüft am 08.05.2022.

Zweig, Katharina A. (2019): Algorithmische Entscheidungen: Transparenz und Kontrolle. Berlin: Konrad-Adenauer-Stiftung (Digitale Gesellschaft, Nr. 338). Online verfügbar unter https://www.kas.de/c/document_library/get_file?uuid=533ef913-e567-987d-54c3-1906395cdb81&groupId=252038, zuletzt geprüft am 24.06.2022.

Online Zeitungsartikel

Allyn, Bobby (2020): America Reckons with Racial Injustice. 'The Computer Got It Wrong': How Facial Recognition Led to False Arrest Of Black Man. In: *NPR* 2020, 24.06.2020. Online verfügbar unter https://www.npr.org/2020/06/24/882683463/the-computer-got-it-wrong-how-facial-recognition-led-to-a-false-arrest-in-michig?t=1654118994957, zuletzt geprüft am 01.06.2022.

Baumhaus, Martin (2016): Das Gold der post-industriellen Gesellschaft. Schürfen Sie Daten wie Gold und passen Sie auf die gefundenen Schätze auf. Denn die Datensammlungen werden richtig wertvoll. Ein Gastbeitrag. In: WirtschaftsWoche, 22.01.2016. Online verfügbar unter https://www.wiwo.de/unternehmen/it/daten-das-gold-der-post-industriellen-gesellschaft-/12844090.html, zuletzt geprüft am 13.05.2022.

Biselli, Anna; Meister, Andre (2019): Asylbehörde sucht mit Künstlicher Intelligenz nach auffälligen Geflüchteten. Bundesamt für Migration und Flüchtlinge: In: *netzpolitik.org* 2019, 19.07.2019 (in Technologie - keine Ergänzungen). Online verfügbar unter https://netzpolitik.org/2019/asylbehoerde-sucht-mit-kuenstlicher-intelligenz-nach-auffaelligen-gefluechteten/, zuletzt geprüft am 20.05.2022.

Brühl, Jannis; Hurtz, Simon (2020): Eine Software schockiert Amerika. Gesichtserkennung mit "Clearview AI". Der Albtraum für die Privatsphäre: Das Programm "Clearview AI" kennt die Gesichter von Millionen Menschen. US-Polizisten setzen sie bereits ein. In: *Süddeutsche Zeitung SZ*, 20.01.2020 (online). Online verfügbar unter https://www.sueddeutsche.de/digital/gesichtserkennung-clearview-app-polizei-gesicht-1.4764389, zuletzt geprüft am 22.05.2022.

Dastin, Jeffrey (2018): Amazon scraps secret AI recruiting tool that showed bias against women. SAN FRANCISCO (Reuters) - Amazon.com Inc's AMZN.O machine-learning specialists uncovered a big problem: their new recruiting engine did not like women. In: *Reuters* OCTOBER 11, 2018, 2018. Online verfügbar unter https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G, zuletzt geprüft am 22.05.2022.

Delcker, Janosch (2019): Europe’s silver bullet in global AI battle: Ethics. EU experts hope ‘trust’ will prove to be the bloc’s competitive edge. In: *POLITICO* 2019, 17.03.2019. Online verfügbar unter https://www.politico.eu/article/europe-silver-bullet-global-ai-battle-ethics/, zuletzt geprüft am 29.05.2022.

Dierks, Benjamin (2017): Die digitalen Tagelöhner. Berufsbild „Crowdworker“. In: *Deutschlandfunk*, 28.09.2017. Online verfügbar unter https://www.deutschlandfunk.de/berufsbild-crowdworker-die-digitalen-tageloehner-100.html, zuletzt geprüft am 01.06.2022.

Forester, Brett (2020): CNN puts itself in the news with ‘Something Else’ label. U.S. network apologizes for ‘poor choice of words’ during election coverage. In: *National News*

2020, 05.11.2020. Online verfügbar unter https://www.aptnnews.ca/national-news/cnn-puts-itself-in-the-news-with-something-else-label/, zuletzt geprüft am 26.05.2022.

Gershgorn, Dave (2017): The data that transformed AI research—and possibly the world. IT'S NOT ABOUT THE ALGORITHM. In: *QUARTZ*, 26.07.2017. Online verfügbar unter https://qz.com/1034972/the-data-that-changed-the-direction-of-ai-research-and-possibly-the-world/, zuletzt geprüft am 05.06.2022.

Waleczek, Torben (2013): Merkels "Neuland" wird zur Lachnummer im Netz. Die Kanzlerin und das Internet. In: *Der Tagesspiegel*, 19.06.2013. Online verfügbar unter https://www.tagesspiegel.de/politik/die-kanzlerin-und-das-internet-merkels-neuland-wird-zur-lachnummer-im-netz/8375974.html, zuletzt geprüft am 18.04.2022.

Ausgewählte Publikationen für SLR

P2: Acuna, Daniel E.; Liang, Lizhen (2021): Are AI Ethics Conferences Different and More Diverse Compared to Traditional Computer Science Conferences? In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery, S. 307–315. Online verfügbar unter https://doi.org/10.1145/3461702.3462616, zuletzt geprüft am 01.07.2022.

P10: Alvero, A. J.; Arthurs, Noah; antonio, anthony lising; Domingue, Benjamin W.; Gebre-Medhin, Ben; Giebel, Sonia; Stevens, Mitchell L. (2020): AI and Holistic Review: Informing Human Reading in College Admissions. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery, S. 200–206. Online verfügbar unter https://doi.org/10.1145/3375627.3375871, zuletzt geprüft am 01.07.2022.

P18: Barocas, Solon; Guo, Anhong; Kamar, Ece; Krones, Jacquelyn; Morris, Meredith Ringel; Vaughan, Jennifer Wortman et al. (2021): Designing Disaggregated Evaluations of AI Systems: Choices, Considerations, and Tradeoffs. In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery, S. 368–378. Online verfügbar unter https://doi.org/10.1145/3461702.3462610, zuletzt geprüft am 01.07.2022.

P23: Benthall, Sebastian; Goldenfein, Jake (2021): Artificial Intelligence and the Purpose of Social Systems. In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery, S. 3–12. Online

verfügbar unter https://doi.org/10.1145/3461702.3462526, zuletzt geprüft am 01.07.2022.

P28: Bondi, Elizabeth; Xu, Lily; Acosta-Navas, Diana; Killian, Jackson A. (2021): Envisioning Communities: A Participatory Approach Towards AI for Social Good. In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery, S. 425–436. Online verfügbar unter https://doi.org/10.1145/3461702.3462612, zuletzt geprüft am 01.07.2022.

P29: Bryant, De'Aira; Howard, Ayanna (2019): A Comparative Analysis of Emotion-Detecting AI Systems with Respect to Algorithm Performance and Dataset Diversity. In: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery (AIES '19), S. 377–382. Online verfügbar unter https://doi.org/10.1145/3306618.3314284, zuletzt geprüft am 01.07.2022.

P31: Kuhlman, Caitlin; Jackson, Latifa; Chunara, Rumi (2020): No Computation without Representation: Avoiding Data and Algorithm Biases through Diversity. In: *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. Online verfügbar unter https://arxiv.org/abs/2002.11836, zuletzt geprüft am 01.07.2022.

P41: Chi, Nicole; Lurie, Emma; Mulligan, Deirdre K. (2021): Reconfiguring Diversity and Inclusion for AI Ethics. In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery, S. 447–457. Online verfügbar unter https://doi.org/10.1145/3461702.3462622, zuletzt geprüft am 01.07.2022.

P45: Cooper, A. Feder; Abrams, Ellen (2021): Emergent Unfairness in Algorithmic Fairness-Accuracy Trade-Off Research. In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery, S. 46–54. Online verfügbar unter https://doi.org/10.1145/3461702.3462519, zuletzt geprüft am 01.07.2022.

P47: Croeser, Sky; Eckersley, Peter (2019): Theories of Parenting and Their Application to Artificial Intelligence. In: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery (AIES '19), S. 423–428. Online verfügbar unter https://doi.org/10.1145/3306618.3314231, zuletzt geprüft am 01.07.2022.

P49: Cruz Cortés, Efrén; Ghosh, Debashis (2020): An Invitation to System-Wide Algorithmic Fairness. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery, S. 235–241. Online verfügbar unter https://doi.org/10.1145/3375627.3375860, zuletzt geprüft am 01.07.2022.

P65: Galdon Clavell, Gemma; Martin Zamorano, Mariano; Castillo, Carlos; Smith, Oliver; Matic, Aleksandar (2020): Auditing Algorithms: On Lessons Learned and the Risks of Data Minimization. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery, S. 265–271. Online verfügbar unter https://doi.org/10.1145/3375627.3375852, zuletzt geprüft am 01.07.2022.

P71: Gilbert, Thomas Krendl; Mintz, Yonatan (2019): Epistemic Therapy for Bias in Automated Decision-Making. In: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery (AIES '19), S. 61–67. Online verfügbar unter https://doi.org/10.1145/3306618.3314294, zuletzt geprüft am 01.07.2022.

P106: Kak, Amba (2020): The Global South is Everywhere, but Also Always Somewhere": National Policy Narratives and AI Justice. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery, S. 307–312. Online verfügbar unter https://doi.org/10.1145/3375627.3375859, zuletzt geprüft am 01.07.2022.

P117: Kshirsagar, Meghana; Robinson, Caleb; Yang, Siyu; Gholami, Shahrzad; Klyuzhin, Ivan; Mukherjee, Sumit et al. (2021): Becoming Good at AI for Good. In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery, S. 664–673. Online verfügbar unter https://doi.org/10.1145/3461702.3462599, zuletzt geprüft am 01.07.2022.

P127: Leavy, Susan; Siapera, Eugenia; O'Sullivan, Barry (2021): Ethical Data Curation for AI: An Approach Based on Feminist Epistemology and Critical Theories of Race. In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery, S. 695–703. Online verfügbar unter https://dl.acm.org/doi/pdf/10.1145/3461702.3462598, zuletzt geprüft am 01.07.2022.

P129: Lee, Michelle Seng Ah; Singh, Jatinder (2021): Risk Identification Questionnaire for Detecting Unintended Bias in the Machine Learning Development Lifecycle. In:

Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery, S. 704–714. Online verfügbar unter https://doi.org/10.1145/3461702.3462572, zuletzt geprüft am 01.07.2022.

P140: Jo, Eun Seo; Gebru, Timnit (2020): Lessons from Archives: Strategies for Collecting Sociocultural Data in Machine Learning. Online verfügbar unter https://arxiv.org/abs/1912.10389, zuletzt geprüft am 01.07.2022.

P165: Park, Joon Sung; Bernstein, Michael S.; Brewer, Robin N.; Kamar, Ece; Morris, Meredith Ringel (2021): Understanding the Representation and Representativeness of Age in AI Data Sets. In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery, S. 834–842. Online verfügbar unter https://arxiv.org/abs/2103.09058, zuletzt geprüft am 01.07.2022.

P170: Posada, Julian; Weller, Nicholas; Wong, Wendy H. (2021): We Haven't Gone Paperless Yet: Why the Printing Press Can Help Us Understand Data and AI. In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery, S. 864–872. Online verfügbar unter https://doi.org/10.1145/3461702.3462604, zuletzt geprüft am 01.07.2022.

P184: Schelenz, Laura; Bison, Ivano; Busso, Matteo; Götzen, Amalia de; Gatica-Perez, Daniel; Giunchiglia, Fausto et al. (2021): The Theory, Practice, and Ethical Challenges of Designing a Diversity-Aware Platform for Social Relations. In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery, S. 905–915. Online verfügbar unter https://doi.org/10.1145/3461702.3462595, zuletzt geprüft am 01.07.2022.

P192: Sharma, Shubham; Zhang, Yunfeng; R\'ıos Aliaga, Jesús M.; Bouneffouf, Djallel; Muthusamy, Vinod; Varshney, Kush R. (2020): Data Augmentation for Discrimination Prevention and Bias Disambiguation. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery, S. 358–364. Online verfügbar unter https://doi.org/10.1145/3375627.3375865, zuletzt geprüft am 01.07.2022.

P204: Leavy, Susan; O'Sullivan, Barry; Siapera, Eugenia (2020): Data, Power and Bias in Artificial Intelligence. Online verfügbar unter https://arxiv.org/abs/2008.07341, zuletzt geprüft am 01.07.2022.

P209: Tommasi, Tatiana; Bucci, Silvia; Caputo, Barbara; Asinari, Pietro (2021): Towards Fairness Certification in Artificial Intelligence. In: *ArXiv* abs/2106.02498. Online verfügbar unter https://arxiv.org/abs/2106.02498, zuletzt geprüft am 01.07.2022.

P212: Tomasev, Nenad; McKee, Kevin R.; Kay, Jackie; Mohamed, Shakir (2021): Fairness for Unobserved Characteristics: Insights from Technological Impacts on Queer Communities. In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery, S. 254–265. Online verfügbar unter https://arxiv.org/pdf/2102.04257, zuletzt geprüft am 01.07.2022.

P219: Vredenburgh, Kate (2021): Alienation in the AI-Driven Workplace. In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery. Online verfügbar unter https://doi.org/10.1145/3461702.3462520, zuletzt geprüft am 01.07.2022.

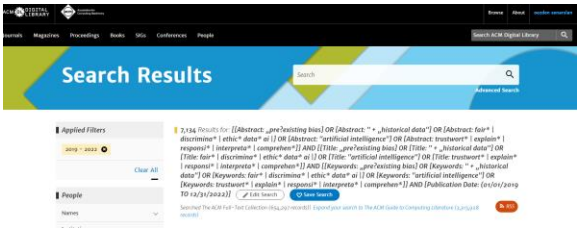
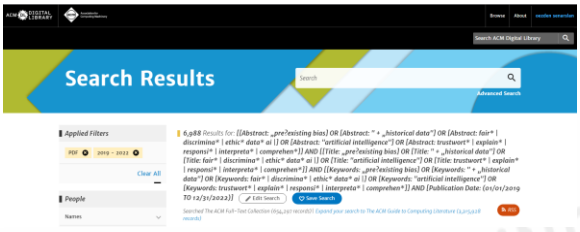
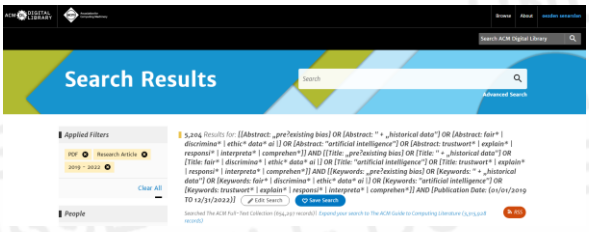
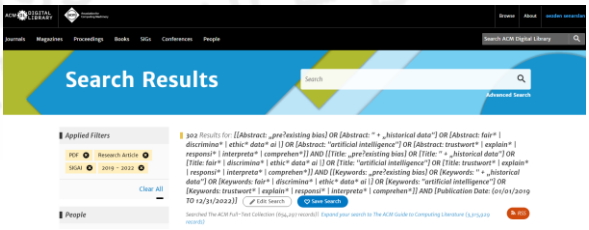
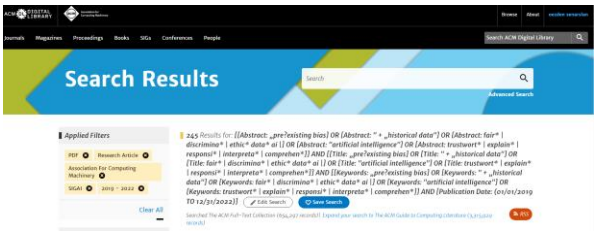
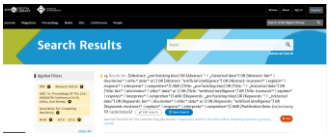
P222: Whittlestone, Jess; Nyrup, Rune; Alexandrova, Anna; Cave, Stephen (2019): The Role and Limits of Principles in AI Ethics: Towards a Focus on Tensions. In: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery (AIES '19), S. 195–200. Online verfügbar unter https://doi.org/10.1145/3306618.3314289, zuletzt geprüft am 01.07.2022.

P238: Zuiderveen Borgesius, Frederik J. (2020): Strengthening legal protection against discrimination by algorithms and artificial intelligence. Online verfügbar unter https://doi.org/10.1080/13642987.2020.1743976, zuletzt geprüft am 01.07.2022.

Anhang

Anhang 1: ACM Digital Library

Tab. I: Selektionsvorgehensweise der Publikationen ACM

Filter	Suche	Treffer
Zeit: 2019-2022		7134
Dateiformat: PDF		6988
Research Article		5204
Special Interest Group on Artificial Intelligence (SIGAI)		302
Association For Computing Machinery		245
Filterung *AIES (AI, Ethics, And Society), Transactions on Knowledge Discovery From Data*	 AIES'21 (74 Treffer)  AIES'19 (39 Treffer)  AIES'20 (43 Treffer)  TKKD (34 Treffer)	190

Literaturangaben zu Suchergebnissen in ACM Digital Library

Abid, Abubakar; Farooqi, Maheen; Zou, James (2021): Persistent Anti-Muslim Bias in Large Language Models. In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery, S. 298–306.

Acuna, Daniel E.; Liang, Lizhen (2021): Are AI Ethics Conferences Different and More Diverse Compared to Traditional Computer Science Conferences? In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery, S. 307–315.

Addison, Arifah; Bartneck, Christoph; Yogeeswaran, Kumar (2019): Robots Can Be More Than Black And White: Examining Racial Bias Towards Robots. In: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery (AIES ’19), S. 493–498.

Afnan, Michael Anis Mihdi; Rudin, Cynthia; Conitzer, Vincent; Savulescu, Julian; Mishra, Abhishek; Liu, Yanhe; Afnan, Masoud (2021): Ethical Implementation of Artificial Intelligence to Select Embryos in In Vitro Fertilization. In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery, S. 316–326.

Aka, Osman; Burke, Ken; Bauerle, Alex; Greer, Christina; Mitchell, Margaret (2021): Measuring Model Biases in the Absence of Ground Truth. In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery, S. 327–335.

Ali, Junaid; Lahoti, Preethi; Gummadi, Krishna P. (2021): Accounting for Model Uncertainty in Algorithmic Discrimination. In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery, S. 336–345.

Ali, Junaid; Zafar, Muhammad Bilal; Singla, Adish; Gummadi, Krishna P. (2019): Loss-Aversively Fair Classification. In: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery (AIES ’19), S. 211–218.

Almeida, Matthew; Zhuang, Yong; Ding, Wei; Crouter, Scott E.; Chen, Ping (2021): Mitigating Class-Boundary Label Uncertainty to Reduce Both Model Bias and Variance. In: ACM Trans. Knowl. Discov. Data 15 (2). DOI: 10.1145/3429447.

Alvero, A. J.; Arthurs, Noah; antonio, anthony lising; Domingue, Benjamin W.; Gebre-Medhin, Ben; Giebel, Sonia; Stevens, Mitchell L. (2020): AI and Holistic Review:

Informing Human Reading in College Admissions. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery, S. 200–206.

Amini, Alexander; Soleimany, Ava P.; Schwarting, Wilko; Bhatia, Sangeeta N.; Rus, Daniela (2019): Uncovering and Mitigating Algorithmic Bias through Learned Latent Structure. In: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery (AIES '19), S. 289–295.

Amornbunchornvej, Chainarong; Berger-Wolf, Tanya (2020): Framework for Inferring Following Strategies from Time Series of Movement Data. In: ACM Trans. Knowl. Discov. Data 14 (3). DOI: 10.1145/3385730.

Amornbunchornvej, Chainarong; Surasvadi, Navaporn; Plangprasopchok, Anon; Thajchayapong, Suttipong (2021): Identifying Linear Models in Multi-Resolution Population Data Using Minimum Description Length Principle to Predict Household Income. In: ACM Trans. Knowl. Discov. Data 15 (2). DOI: 10.1145/3424670.

Avin, Shahar; Gruetzemacher, Ross; Fox, James (2020): Exploring AI Futures Through Role Play. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery, S. 8–14.

Azevedo, Ricardo de; Machado, Gabriel Resende; Goldschmidt, Ronaldo Ribeiro; Choren, Ricardo (2020): A Reduced Network Traffic Method for IoT Data Clustering. In: ACM Trans. Knowl. Discov. Data 15 (1). DOI: 10.1145/3423139.

Bakker, Michiel A.; Tu, Duy Patrick; Gummadi, Krishna P.; Pentland, Alex Sandy; Varshney, Kush R.; Weller, Adrian (2021): Beyond Reasonable Doubt: Improving Fairness in Budget-Constrained Decision Making Using Confidence Thresholds. In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery, S. 346–356.

Barocas, Solon; Guo, Anhong; Kamar, Ece; Krones, Jacquelyn; Morris, Meredith Ringel; Vaughan, Jennifer Wortman et al. (2021): Designing Disaggregated Evaluations of AI Systems: Choices, Considerations, and Tradeoffs. In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery, S. 368–378.

Behzadan, Vahid; Minton, James; Munir, Arslan (2019): TrolleyMod v1.0: An Open-Source Simulation and Data-Collection Platform for Ethical Decision Making in Autonomous Vehicles. In: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery (AIES '19), S. 391–395.

Belfield, Haydn (2020): Activism by the AI Community: Analysing Recent Achievements and Future Prospects. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery, S. 15–21.

Belitz, Clara; Jiang, Lan; Bosch, Nigel (2021): Automating Procedurally Fair Feature Selection in Machine Learning. In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery, S. 379–389.

Ben David, Daniel; Resheff, Yehezkel S.; Tron, Talia (2021): Explainable AI and Adoption of Financial Algorithmic Advisors: An Experimental Study. In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery, S. 390–400.

Benthall, Sebastian; Goldenfein, Jake (2021): Artificial Intelligence and the Purpose of Social Systems. In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery, S. 3–12.

Beutel, Alex; Chen, Jilin; Doshi, Tulsee; Qian, Hai; Woodruff, Allison; Luu, Christine et al. (2019): Putting Fairness Principles into Practice: Challenges, Metrics, and Improvements. In: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery (AIES '19), S. 453–459.

Bian, Jiang; Xiong, Haoyi; Fu, Yanjie; Huan, Jun; Guo, Zhishan (2020): MP2SDA: Multi-Party Parallelized Sparse Discriminant Learning. In: ACM Trans. Knowl. Discov. Data 14 (3). DOI: 10.1145/3374919.

Biswas, Arpita; Mukherjee, Suvam (2021): Ensuring Fairness under Prior Probability Shifts. In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery, S. 414–424.

Bondi, Elizabeth; Xu, Lily; Acosta-Navas, Diana; Killian, Jackson A. (2021): Envisioning Communities: A Participatory Approach Towards AI for Social Good. In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery, S. 425–436.

Bryant, De'Aira; Howard, Ayanna (2019): A Comparative Analysis of Emotion-Detecting AI Systems with Respect to Algorithm Performance and Dataset Diversity. In: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery (AIES '19), S. 377–382.

Cai, William; Gaebler, Johann; Garg, Nikhil; Goel, Sharad (2020): Fair Allocation through Selective Information Acquisition. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery, S. 22–28.

Cave, Stephen (2020): The Problem with Intelligence: Its Value-Laden History and the Future of AI. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery, S. 29–35.

Cave, Stephen; Coughlan, Kate; Dihal, Kanta (2019): Scary Robots": Examining Public Responses to AI. In: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery (AIES ’19), S. 331–337.

Chakraborti, Tathagata; Kambhampati, Subbarao (2019): (When) Can AI Bots Lie? In: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery (AIES ’19), S. 53–59.

Chan, Lok; Doyle, Kenzie; McElfresh, Duncan; Conitzer, Vincent; Dickerson, John P.; Schaich Borg, Jana; Sinnott-Armstrong, Walter (2020): Artificial Artificial Intelligence: Measuring Influence of AI ’Assessments’ on Moral Decision-Making. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery, S. 214–220.

Chaput, Rémy; Duval, Jérémy; Boissier, Olivier; Guillermin, Mathieu; Hassas, Salima (2021): A Multi-Agent Approach to Combine Reasoning and Learning for an Ethical Behavior. In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery, S. 13–23.

Chen, Violet; Hooker, J. N. (2020): A Just Approach Balancing Rawlsian Leximax Fairness and Utilitarianism. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery, S. 221–227.

Chen, Yan; Mahoney, Christopher; Grasso, Isabella; Wali, Esma; Matthews, Abigail; Middleton, Thomas et al. (2021): Gender Bias and Under-Representation in Natural Language Processing Across Human Languages. In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery, S. 24–34.

Chi, Nicole; Lurie, Emma; Mulligan, Deirdre K. (2021): Reconfiguring Diversity and Inclusion for AI Ethics. In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery, S. 447–457.

Chuan, Ching-Hua; Tsai, Wan-Hsiu Sunny; Cho, Su Yeon (2019): Framing Artificial Intelligence in American Newspapers. In: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery (AIES ’19), S. 339–344.

Cihon, Peter; Maas, Matthijs M.; Kemp, Luke (2020): Should Artificial Intelligence Governance Be Centralised? Design Lessons from History. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery, S. 228–234.

Cooper, A. Feder; Abrams, Ellen; NA, N. A. (2021): Emergent Unfairness in Algorithmic Fairness-Accuracy Trade-Off Research. In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery, S. 46–54.

Coston, Amanda; Ramamurthy, Karthikeyan Natesan; Wei, Dennis; Varshney, Kush R.; Speakman, Skyler; Mustahsan, Zairah; Chakraborty, Supriyo (2019): Fair Transfer Learning with Missing Protected Attributes. In: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery (AIES ’19), S. 91–98.

Croeser, Sky; Eckersley, Peter (2019): Theories of Parenting and Their Application to Artificial Intelligence. In: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery (AIES ’19), S. 423–428.

Cruz, Joe (2019): Shared Moral Foundations of Embodied Artificial Intelligence. In: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery (AIES ’19), S. 139–146.

Cruz Cortés, Efrén; Ghosh, Debashis (2020): An Invitation to System-Wide Algorithmic Fairness. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery, S. 235–241.

Dai, Jessica; Fazelpour, Sina; Lipton, Zachary (2021): Fair Machine Learning Under Partial Compliance. In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery, S. 55–65.

Daniele, Antonio; Song, Yi-Zhe (2019): AI + Art = Human. In: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery (AIES ’19), S. 155–161.

Diana, Emily; Gill, Wesley; Kearns, Michael; Kenthapadi, Krishnaram; Roth, Aaron (2021): Minimax Group Fairness: Algorithms and Experiments. In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery, S. 66–76.

Duckworth, Paul; Graham, Logan; Osborne, Michael (2019): Inferring Work Task Automatability from AI Expert Evidence. In: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery (AIES '19), S. 485–491.

Ema, Arisa; Nagakura, Katsue; Fujita, Takanori (2020): Proposal for Type Classification for Building Trust in Medical Artificial Intelligence Systems. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery, S. 251–257.

Erdélyi, Olivia J.; Erdélyi, Gábor (2020): The AI Liability Puzzle and a Fund-Based Work-Around. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery, S. 50–56.

Ermiş, Beyza; Cemgundefinedl, A. Taylan (2020): Data Sharing via Differentially Private Coupled Matrix Factorization. In: ACM Trans. Knowl. Discov. Data 14 (3). DOI: 10.1145/3372408.

Fazelpour, Sina; Lipton, Zachary C. (2020): Algorithmic Fairness from a Non-Ideal Perspective. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery, S. 57–63.

Fernandes, Pedro M.; Santos, Francisco C.; Lopes, Manuel (2020): Adoption Dynamics and Societal Impact of AI Systems in Complex Networks. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery, S. 258–264.

Fish, Benjamin; Stark, Luke (2021): Reflexive Design for Fairness and Other Human Values in Formal Models. In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery, S. 89–99.

Flathmann, Christopher; Schelble, Beau G.; Zhang, Rui; McNeese, Nathan J. (2021): Modeling and Guiding the Creation of Ethical Human-AI Teams. In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery, S. 469–479.

Fleisher, Will (2021): What’s Fair about Individual Fairness? In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery, S. 480–490.

Galdon Clavell, Gemma; Mart\’ın Zamorano, Mariano; Castillo, Carlos; Smith, Oliver; Matic, Aleksandar (2020): Auditing Algorithms: On Lessons Learned and the Risks of Data Minimization. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery, S. 265–271.

Galhotra, Sainyam; Saisubramanian, Sandhya; Zilberstein, Shlomo (2021): Learning to Generate Fair Clusters from Demonstrations. In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery, S. 491–501.

Garg, Sahaj; Perot, Vincent; Limtiaco, Nicole; Taly, Ankur; Chi, Ed H.; Beutel, Alex (2019): Counterfactual Fairness in Text Classification through Robustness. In: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery (AIES ’19), S. 219–226.

Garrett, Natalie; Beard, Nathan; Fiesler, Casey (2020): More Than "If Time Allows": The Role of Ethics in AI Education. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery, S. 272–278.

Ghosh, Bishwamittra; Meel, Kuldeep S. (2019): IMLI: An Incremental Framework for MaxSAT-Based Learning of Interpretable Classification Rules. In: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery (AIES ’19), S. 203–210.

Gilbert, Thomas Krendl; Mintz, Yonatan (2019): Epistemic Therapy for Bias in Automated Decision-Making. In: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery (AIES ’19), S. 61–67.

Goel, Naman; Faltings, Boi (2019): Crowdsourcing with Fairness, Diversity and Budget Constraints. In: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery (AIES ’19), S. 297–304.

Goel, Kanika; Leemans, Sander J. J.; Martin, Niels; Wynn, Moe T. (2022): Quality-Informed Process Mining: A Case for Standardised Data Quality Annotations. In: ACM Trans. Knowl. Discov. Data 16 (5). DOI: 10.1145/3511707.

Golden, Paige; Danks, David (2021): Ethical Obligations to Provide Novelty. In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery, S. 502–508.

Goodman, Bryce (2021): Hard Choices and Hard Limits in Artificial Intelligence. In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery, S. 112–121.

Gram-Hansen, Bradley J.; Helber, Patrick; Varatharajan, Indhu; Azam, Faiza; Coca-Castro, Alejandro; Kopackova, Veronika; Bilinski, Piotr (2019): Mapping Informal Settlements in Developing Countries Using Machine Learning and Low Resolution Multi-Spectral Data. In: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery (AIES '19), S. 361–368.

Green, Nancy (2021): An AI Ethics Course Highlighting Explicit Ethical Agents. In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery, S. 519–524.

Guidotti, Riccardo; Monreale, Anna (2021): Designing Shapelets for Interpretable Data-Agnostic Classification. In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery, S. 532–542.

Guo, Wei; Caliskan, Aylin (2021): Detecting Emergent Intersectional Biases: Contextualized Word Embeddings Contain a Distribution of Human-like Biases. In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery, S. 122–133.

Guo, Yuan; Sun, Yu; Wu, Kai; Jiang, Kerong (2020): New Algorithms of Feature Selection and Big Data Assignment for CBR System Integrated by Bayesian Network. In: ACM Trans. Knowl. Discov. Data 14 (2). DOI: 10.1145/3373086.

Guo, Jipeng; Sun, Yanfeng; Gao, Junbin; Hu, Yongli; Yin, Baocai (2020): Robust Adaptive Linear Discriminant Analysis with Bidirectional Reconstruction Constraint. In: ACM Trans. Knowl. Discov. Data 14 (6). DOI: 10.1145/3409478.

Guo, Mengzhuo; Xu, Zhongzhi; Zhang, Qingpeng; Liao, Xiuwu; Liu, Jiapeng (2021): Deciphering Feature Effects on Decision-Making in Ordinal Regression Problems: An Explainable Ordinal Factorization Model. In: ACM Trans. Knowl. Discov. Data 16 (3). DOI: 10.1145/3487048.

Hadfield-Menell, Dylan; Andrus, Mckane; Hadfield, Gillian (2019): Legible Normativity for AI Alignment: The Value of Silly Rules. In: Proceedings of the 2019

AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery (AIES ’19), S. 115–121.

Han, The Anh; Pereira, Lu\’ıs Moniz; Lenaerts, Tom (2019): Modelling and Influencing the AI Bidding War: A Research Agenda. In: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery (AIES ’19), S. 5–11.

Hannan, Jacqueline; Chen, Huei-Yen Winnie; Joseph, Kenneth (2021): Who Gets What, According to Whom? An Analysis of Fairness Perceptions in Service Allocation. In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery, S. 555–565.

He, Yuzi; Burghardt, Keith; Lerman, Kristina (2020): A Geometric Solution to Fair Representations. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery, S. 279–285.

Henriksen, Anne; Enni, Simon; Bechmann, Anja (2021): Situated Accountability: Ethical Principles, Certification Standards, and Explanation Methods in Applied AI. In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery, S. 574–585.

Henzgen, Sascha; Hüllermeier, Eyke (2019): Mining Rank Data. In: ACM Trans. Knowl. Discov. Data 13 (6). DOI: 10.1145/3363572.

Herington, Jonathan (2020): Measuring Fairness in an Unfair World. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery, S. 286–292.

Hernández-Orallo, José; Vold, Karina (2019): AI Extenders: The Ethical and Societal Implications of Humans Cognitively Extended by AI. In: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery (AIES ’19), S. 507–513.

Hidalgo, Juan I. G.; Santos, Silas G. T. C.; Barros, Roberto S. M. (2021): Dynamically Adjusting Diversity in Ensembles for the Classification of Data Streams with Concept Drift. In: ACM Trans. Knowl. Discov. Data 16 (2). DOI: 10.1145/3466616.

Hind, Michael; Wei, Dennis; Campbell, Murray; Codella, Noel C. F.; Dhurandhar, Amit; Mojsilović, Aleksandra et al. (2019): TED: Teaching AI to Explain Its Decisions. In: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery (AIES ’19), S. 123–129.

Huang, Lingxiao; Wei, Julia; Celis, Elisa (2020): Towards Just, Fair and Interpretable Methods for Judicial Subset Selection. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery, S. 293–299.

Huang, Yourong; Xiao, Zhu; Yu, Xiaoyou; Wang, Dong; Havyarimana, Vincent; Bai, Jing (2019): Road Network Construction with Complex Intersections Based on Sparsely Sampled Private Car Trajectory Data. In: ACM Trans. Knowl. Discov. Data 13 (3). DOI: 10.1145/3326060.

Ibrahim, Mark; Louie, Melissa; Modarres, Ceena; Paisley, John (2019): Global Explanations of Neural Networks: Mapping the Landscape of Predictions. In: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery (AIES ’19), S. 279–287.

Islam, Rashidul; Pan, Shimei; Foulds, James R. (2021): Can We Obtain Fairness For Free? In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery, S. 586–596.

Javadi, Seyyed Ahmad; Cloete, Richard; Cobbe, Jennifer; Lee, Michelle Seng Ah; Singh, Jatinder (2020): Monitoring Misuse for Accountable ’Artificial Intelligence as a Service’. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery, S. 300–306.

Javadi, Seyyed Ahmad; Norval, Chris; Cloete, Richard; Singh, Jatinder (2021): Monitoring AI Services for Misuse. In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery, S. 597–607.

Jiang, Weijie; Pardos, Zachary A. (2021): Towards Equity and Algorithmic Fairness in Student Grade Prediction. In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery, S. 608–617.

Jo, Eun Seo; Gebru, Timnit (2020): Lessons from Archives: Strategies for Collecting So-ciocultural Data in Machine Learning.

Kak, Amba (2020): The Global South is Everywhere, but Also Always Somewhere": National Policy Narratives and AI Justice. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery, S. 307–312.

Karpati, Daniel; Najjar, Amro; Ambrossio, Diego Agustin (2020): Ethics of Food Recommender Applications. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery, S. 313–319.

Kasirzadeh, Atoosa; Clifford, Damian (2021): Fairness and Data Protection Impact Assessments. In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery, S. 146–153.

Kasirzadeh, Atoosa; Klein, Colin (2021): The Ethical Gravity Thesis: Marrian Levels and the Persistence of Bias in Automated Decision-Making Systems. In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery, S. 618–626.

Katib, Anas; Rao, Praveen; Barnard, Kobus; Kamhoua, Charles (2019): Fast Approximate Score Computation on Large-Scale Distributed Data for Learning Multinomial Bayesian Networks. In: ACM Trans. Knowl. Discov. Data 13 (2). DOI: 10.1145/3301304.

Kelley, Patrick Gage; Yang, Yongwei; Heldreth, Courtney; Moessner, Christopher; Sedley, Aaron; Kramm, Andreas et al. (2021): Exciting, Useful, Worrying, Futuristic: Public Perception of Artificial Intelligence in 8 Countries. In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery, S. 627–637.

Kim, Eugenia; Bryant, De'Aira; Srikanth, Deepak; Howard, Ayanna (2021): Age Bias in Emotion Detection: An Analysis of Facial Emotion Recognition Performance on Young, Middle-Aged, and Older Adults. In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery, S. 638–644.

Kim, Michael P.; Ghorbani, Amirata; Zou, James (2019): Multiaccuracy: Black-Box Post-Processing for Fairness in Classification. In: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery (AIES '19), S. 247–254.

Kommiya Mothilal, Ramaravind; Mahajan, Divyat; Tan, Chenhao; Sharma, Amit (2021): Towards Unifying Feature Attribution and Counterfactual Explanations: Different Means to the Same End. In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery, S. 652–663.

Krafft, P. M.; Young, Meg; Katell, Michael; Huang, Karen; Bugingo, Ghislain (2020): Defining AI in Policy versus Practice. In: Proceedings of the AAAI/ACM

Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery, S. 72–78.

Kshirsagar, Meghana; Robinson, Caleb; Yang, Siyu; Gholami, Shahrzad; Klyuzhin, Ivan; Mukherjee, Sumit et al. (2021): Becoming Good at AI for Good. In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery, S. 664–673.

Kuhlman, Caitlin; Gerych, Walter; Rundensteiner, Elke (2021): Measuring Group Advantage: A Comparative Study of Fair Ranking Metrics. In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery, S. 674–682.

L. Cardoso, Rodrigo; Meira Jr., Wagner; Almeida, Virgilio; J. Zaki, Mohammed (2019): A Framework for Benchmarking Discrimination-Aware Models in Machine Learning. In: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery (AIES ’19), S. 437–444.

Lakkaraju, Himabindu; Bastani, Osbert (2020): How Do I Fool You?": Manipulating User Trust via Misleading Black Box Explanations. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery, S. 79–85.

Lakkaraju, Himabindu; Kamar, Ece; Caruana, Rich; Leskovec, Jure (2019): Faithful and Customizable Explanations of Black Box Models. In: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery (AIES ’19), S. 131–138.

Lappas, Theodoros (2020): Mining Career Paths from Large Resume Databases: Evidence from IT Professionals. In: ACM Trans. Knowl. Discov. Data 14 (3). DOI: 10.1145/3379984.

Larsen, Benjamin Cedric (2021): A Framework for Understanding AI-Induced Field Change: How AI Technologies Are Legitimized and Institutionalized. In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery, S. 683–694.

Leavy, Susan; Siapera, Eugenia; O’Sullivan, Barry (2021): Ethical Data Curation for AI: An Approach Based on Feminist Epistemology and Critical Theories of Race. In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery, S. 695–703.

Leben, Derek (2020): Normative Principles for Evaluating Fairness in Machine Learning. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery, S. 86–92.

Lee, Michelle Seng Ah; Singh, Jatinder (2021): Risk Identification Questionnaire for Detecting Unintended Bias in the Machine Learning Development Lifecycle. In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery, S. 704–714.

Li, Lan; Lassiter, Tina; Oh, Joohee; Lee, Min Kyung (2021): Algorithmic Hiring in Practice: Recruiter and HR Professional's Perspectives on AI Use in Hiring. In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery, S. 166–176.

Li, Yanni; Liu, Bing; Yu, Yongbo; Li, Hui; Sun, Jiacan; Cui, Jiangtao (2021): 3E-LDA: Three Enhancements to Linear Discriminant Analysis. In: ACM Trans. Knowl. Discov. Data 15 (4). DOI: 10.1145/3442347.

Li, Qingyang; Yu, Zhiwen; Guo, Bin; Xu, Huang; Lu, Xinjiang (2019): Housing Demand Estimation Based on Express Delivery Data. In: ACM Trans. Knowl. Discov. Data 13 (4). DOI: 10.1145/3332522.

Liu, Hao; Guo, Qingyu; Zhu, Hengshu; Zhuang, Fuzhen; Yang, Shenwen; Dou, Dejing; Xiong, Hui (2022): Who Will Win the Data Science Competition? Insights from KDD Cup 2019 and Beyond. In: ACM Trans. Knowl. Discov. Data 16 (5). DOI: 10.1145/3511896.

Liu, David; Shafi, Zohair; Fleisher, William; Eliassi-Rad, Tina; Alfeld, Scott (2021): RAWLSNET: Altering Bayesian Networks to Encode Rawlsian Fair Equality of Opportunity. In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery, S. 745–755.

Liu, Bo; Zhong, Haowen; Xiao, Yanshan (2021): New Multi-View Classification Method with Uncertain Data. In: ACM Trans. Knowl. Discov. Data 16 (1). DOI: 10.1145/3458282.

Loi, Michele; Herlitz, Anders; Heidari, Hoda (2021): Fair Equality of Chances for Prediction-Based Decisions. In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery, S. 756.

Loi, Michele; Spielkamp, Matthias (2021): Towards Accountability in the Use of Artificial Intelligence for Public Administrations. In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery, S. 757–766.

Maitra, Suvradip (2020): Artificial Intelligence and Indigenous Perspectives: Protecting and Empowering Intelligent Human Beings. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery, S. 320–326.

Mart\'ınez-Plumed, Fernando; Tolan, Songül; Pesole, Annarosa; Hernández-Orallo, José; Fernández-Mac\'ıas, Enrique; Gómez, Emilia (2020): Does AI Qualify for the Job? A Bidirectional Model Mapping Labour and AI Intensities. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery, S. 94–100.

Mazijn, Carmen; Danckaert, Jan; Ginis, Vincent (2021): How Do the Score Distributions of Subpopulations Influence Fairness Notions? In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery, S. 767–776.

McKee, Heidi A.; Porter, James E. (2020): Ethics for AI Writing: The Importance of Rhetorical Context. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery, S. 110–116.

McNamara, Daniel; Ong, Cheng Soon; Williamson, Robert C. (2019): Costs and Benefits of Fair Representation Learning. In: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery (AIES '19), S. 263–270.

Mhasawade, Vishwali; Chunara, Rumi (2021): Causal Multi-Level Fairness. In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery, S. 784–794.

Mohammadi, Kiarash; Karimi, Amir-Hossein; Barthe, Gilles; Valera, Isabel (2021): Scaling Guarantees for Nearest Counterfactual Explanations. In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery, S. 177–187.

Morgan, Andrew; Pass, Rafael (2019): Paradoxes in Fair Computer-Aided Decision Making. In: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery (AIES '19), S. 85–90.

Nanayakkara, Priyanka; Hullman, Jessica; Diakopoulos, Nicholas (2021): Unpacking the Expressed Consequences of AI Research in Broader Impact Statements. In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery, S. 795–806.

Nashed, Samer; Svegliato, Justin; Zilberstein, Shlomo (2021): Ethically Compliant Planning within Moral Communities. In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery, S. 188–198.

Nie, Feiping; Wang, Zheng; Wang, Rong; Wang, Zhen; Li, Xuelong (2020): Adaptive Local Linear Discriminant Analysis. In: ACM Trans. Knowl. Discov. Data 14 (1). DOI: 10.1145/3369870.

Nielsen, Aileen (2021): Measuring Lay Reactions to Personal Data Markets. In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery, S. 807–813.

Noriega-Campero, Alejandro; Bakker, Michiel A.; Garcia-Bulle, Bernardo; Pentland, Alex 'Sandy' (2019): Active Fairness in Algorithmic Decision Making. In: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery (AIES '19), S. 77–83.

Oneto, Luca; Doninini, Michele; Elders, Amon; Pontil, Massimiliano (2019): Taking Advantage of Multitask Learning for Fair Classification. In: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery (AIES '19), S. 227–237.

Pagnucco, Maurice; Rajaratnam, David; Limarga, Raynaldio; Nayak, Abhaya; Song, Yang (2021): Epistemic Reasoning for Machine Ethics with Situation Calculus. In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery, S. 814–821.

Pandey, Akshat; Caliskan, Aylin (2021): Disparate Impact of Artificial Intelligence Bias in Ridehailing Economy's Price Discrimination Algorithms. In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery, S. 822–833.

Pandya, Ravi; Huang, Sandy H.; Hadfield-Menell, Dylan; Dragan, Anca D. (2019): Human-AI Learning Performance in Multi-Armed Bandits. In: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery (AIES '19), S. 369–375.

Pang, Guansong; Cao, Longbing (2020): Heterogeneous Univariate Outlier Ensembles in Multidimensional Data. In: ACM Trans. Knowl. Discov. Data 14 (6). DOI: 10.1145/3403934.

Park, Joon Sung; Bernstein, Michael S.; Brewer, Robin N.; Kamar, Ece; Morris, Meredith Ringel (2021): Understanding the Representation and Representativeness of

Age in AI Data Sets. In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery, S. 834–842.

Perrier, Elija (2021): Quantum Fair Machine Learning. In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery, S. 843–853.

Perrone, Valerio; Donini, Michele; Zafar, Muhammad Bilal; Schmucker, Robin; Kenthapadi, Krishnaram; Archambeau, Cédric (2021): Fair Bayesian Optimization. In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery, S. 854–863.

Pfohl, Stephen; Marafino, Ben; Coulet, Adrien; Rodriguez, Fatima; Palaniappan, Latha; Shah, Nigam H. (2019): Creating Fair Models of Atherosclerotic Cardiovascular Disease Risk. In: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery (AIES ’19), S. 271–278.

Posada, Julian; Weller, Nicholas; Wong, Wendy H. (2021): We Haven’t Gone Paperless Yet: Why the Printing Press Can Help Us Understand Data and AI. In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery, S. 864–872.

Poyiadzi, Rafael; Sokol, Kacper; Santos-Rodriguez, Raul; Bie, Tijl de; Flach, Peter (2020): FACE: Feasible and Actionable Counterfactual Explanations. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery, S. 344–350.

Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (2019). New York, NY, USA: Association for Computing Machinery (AIES ’19).

Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (2021). New York, NY, USA: Association for Computing Machinery.

Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (2021). New York, NY, USA: Association for Computing Machinery.

Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (2020). New York, NY, USA: Association for Computing Machinery.

Prost, Flavien; Awasthi, Pranjal; Blumm, Nick; Kumthekar, Aditee; Potter, Trevor; Wei, Li et al. (2021): Measuring Model Fairness under Noisy Covariates: A Theoretical Perspective. In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery, S. 873–883.

Prunkl, Carina; Whittlestone, Jess (2020): Beyond Near- and Long-Term: Towards a Clearer Account of Research Priorities in AI Ethics and Society. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery, S. 138–143.

Raji, Inioluwa Deborah; Buolamwini, Joy (2019): Actionable Auditing: Investigating the Impact of Publicly Naming Biased Performance Results of Commercial AI Products. In: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery (AIES '19), S. 429–435.

Raji, Inioluwa Deborah; Gebru, Timnit; Mitchell, Margaret; Buolamwini, Joy; Lee, Joonseok; Denton, Emily (2020): Saving Face: Investigating the Ethical Concerns of Facial Recognition Auditing. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery, S. 145–151.

Roseberry, Martha; Krawczyk, Bartosz; Cano, Alberto (2019): Multi-Label Punitive KNN with Self-Adjusting Memory for Drifting Data Streams. In: ACM Trans. Knowl. Discov. Data 13 (6). DOI: 10.1145/3363573.

Saisubramanian, Sandhya; Galhotra, Sainyam; Zilberstein, Shlomo (2020): Balancing the Tradeoff Between Clustering Value and Interpretability. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery, S. 351–357.

Savva, Fotis; Anagnostopoulos, Christos; Triantafillou, Peter; Kolomvatsos, Kostas (2020): Large-Scale Data Exploration Using Explanatory Regression Functions. In: ACM Trans. Knowl. Discov. Data 14 (6). DOI: 10.1145/3410448.

Saxena, Nripsuta Ani; Huang, Karen; DeFilippis, Evan; Radanovic, Goran; Parkes, David C.; Liu, Yang (2019): How Do Fairness Definitions Fare? Examining Public Attitudes Towards Algorithmic Definitions of Fairness. In: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery (AIES '19), S. 99–106.

Schelenz, Laura; Bison, Ivano; Busso, Matteo; Götzen, Amalia de; Gatica-Perez, Daniel; Giunchiglia, Fausto et al. (2021): The Theory, Practice, and Ethical Challenges of Designing a Diversity-Aware Platform for Social Relations. In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery, S. 905–915.

Schiff, Daniel; Biddle, Justin; Borenstein, Jason; Laas, Kelly (2020): What's Next for AI Ethics, Policy, and Governance? A Global Overview. In: Proceedings of the

AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery, S. 153–158.

Schumann, Candice; Ricco, Susanna; Prabhu, Utsav; Ferrari, Vittorio; Pantofaru, Caroline (2021): A Step Toward More Inclusive People Annotations for Fairness. In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery, S. 916–925.

Schutzman, Zachary (2020): Trade-Offs in Fair Redistricting. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery, S. 159–165.

Segal, Shahar; Adi, Yossi; Pinkas, Benny; Baum, Carsten; Ganesh, Chaya; Keshet, Joseph (2021): Fairness in the Eyes of the Data: Certifying Machine-Learning Models. In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery, S. 926–935.

Shah, Kulin; Gupta, Pooja; Deshpande, Amit; Bhattacharyya, Chiranjib (2021): Rawlsian Fair Adaptation of Deep Learning Classifiers. In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery, S. 936–945.

Sharma, Shubham; Gee, Alan H.; Paydarfar, David; Ghosh, Joydeep (2021): FaiR-N: Fair and Robust Neural Networks for Structured Data. In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery, S. 946–955.

Sharma, Shubham; Henderson, Jette; Ghosh, Joydeep (2020): CERTIFAI: A Common Framework to Provide Explanations and Analyse the Fairness and Robustness of Black-Box Models. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery, S. 166–172.

Sharma, Ms Promila; Meena, Uma; Sharma, Girish Kumar (2022): Intelligent Data Analysis Using Optimized Support Vector Machine Based Data Mining Approach for Tourism Industry. In: ACM Trans. Knowl. Discov. Data 16 (5). DOI: 10.1145/3494566.

Sharma, Shubham; Zhang, Yunfeng; R\'ıos Aliaga, Jesús M.; Bouneffouf, Djallel; Muthusamy, Vinod; Varshney, Kush R. (2020): Data Augmentation for Discrimination Prevention and Bias Disambiguation. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery, S. 358–364.

Shekhar, Shubhranshu; Shah, Neil; Akoglu, Leman (2021): FairOD: Fairness-Aware Outlier Detection. In: Proceedings of the 2021 AAAI/ACM Conference on AI,

Ethics, and Society. New York, NY, USA: Association for Computing Machinery, S. 210–220.

Shevlane, Toby; Dafoe, Allan (2020): The Offense-Defense Balance of Scientific Knowledge: Does Publishing AI Research Reduce Misuse? In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery, S. 173–179.

Shi, Tian; Zhang, Xuchao; Wang, Ping; Reddy, Chandan K. (2021): Corpus-Level and Concept-Based Explanations for Interpretable Document Classification. In: ACM Trans. Knowl. Discov. Data 16 (3). DOI: 10.1145/3477539.

Shokri, Reza; Strobel, Martin; Zick, Yair (2021): On the Privacy Risks of Model Explanations. In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery, S. 231–241.

Shulman, Eyal; Wolf, Lior (2020): Meta Decision Trees for Explainable Recommendation Systems. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery, S. 365–371.

Siddiqui, Md Amran; Fern, Alan; Dietterich, Thomas G.; Wong, Weng-Keen (2019): Sequential Feature Explanations for Anomaly Detection. In: ACM Trans. Knowl. Discov. Data 13 (1). DOI: 10.1145/3230666.

Slack, Dylan; Hilgard, Sophie; Jia, Emily; Singh, Sameer; Lakkaraju, Himabindu (2020): Fooling LIME and SHAP: Adversarial Attacks on Post Hoc Explanation Methods. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery, S. 180–186.

Steinbuss, Georg; Böhm, Klemens (2021): Benchmarking Unsupervised Outlier Detection with Realistic Synthetic Data. In: ACM Trans. Knowl. Discov. Data 15 (4). DOI: 10.1145/3441453.

Sühr, Tom; Hilgard, Sophie; Lakkaraju, Himabindu (2021): Does Fair Ranking Improve Minority Outcomes? Understanding the Interplay of Human and Algorithmic Biases in Online Hiring. In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery, S. 989–999.

Susser, Daniel (2019): Invisible Influence: Artificial Intelligence and the Ethics of Adaptive Choice Architectures. In: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery (AIES '19), S. 403–408.

Susser, Daniel; Grimaldi, Vincent (2021): Measuring Automated Influence: Between Empirical Evidence and Ethical Values. In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery, S. 242–253.

Swinger, Nathaniel; De-Arteaga, Maria; Heffernan IV, Neil Thomas; Leiserson, Mark D. M.; Kalai, Adam Tauman (2019): What Are the Biases in My Word Embedding? In: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery (AIES '19), S. 305–311.

Teso, Stefano; Kersting, Kristian (2019): Explanatory Interactive Machine Learning. In: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery (AIES '19), S. 239–245.

Tomasev, Nenad; McKee, Kevin R.; Kay, Jackie; Mohamed, Shakir (2021): Fairness for Unobserved Characteristics: Insights from Technological Impacts on Queer Communities. In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery, S. 254–265.

Tucker, Aaron D.; Anderljung, Markus; Dafoe, Allan (2020): Social and Governance Implications of Improved Data Efficiency. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery, S. 378–384.

Tuomo, Alasalmi; Suutala, Jaakko; Röning, Juha; Koskimäki, Heli (2020): Better Classifier Calibration for Small Datasets. In: ACM Trans. Knowl. Discov. Data 14 (3). DOI: 10.1145/3385656.

Varde, Aparna S. (2022): Computational Estimation by Scientific Data Mining with Classical Methods to Automate Learning Strategies of Scientists. In: ACM Trans. Knowl. Discov. Data 16 (5). DOI: 10.1145/3502736.

Vredenburgh, Kate (2021): Alienation in the AI-Driven Workplace. In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery, S. 266.

Wang, Yuandong; Lin, Xuelian; Wei, Hua; Wo, Tianyu; Huang, Zhou; Zhang, Yong; Xu, Jie (2019): A Unified Framework with Multi-Source Data for Predicting Passenger Demands of Ride Services. In: ACM Trans. Knowl. Discov. Data 13 (6). DOI: 10.1145/3355563.

Whittlestone, Jess; Nyrup, Rune; Alexandrova, Anna; Cave, Stephen (2019): The Role and Limits of Principles in AI Ethics: Towards a Focus on Tensions. In: Proceedings

of the 2019 AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery (AIES '19), S. 195–200.

Xu, Yanan; Shen, Yanyan; Zhu, Yanmin; Yu, Jiadi (2020): AR2Net: An Attentive Neural Approach for Business Location Selection with Satellite Data and Urban Data. In: ACM Trans. Knowl. Discov. Data 14 (2). DOI: 10.1145/3372406.

Yaghini, Mohammad; Krause, Andreas; Heidari, Hoda (2021): A Human-in-the-Loop Framework to Construct Context-Aware Mathematical Notions of Outcome Fairness. In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery, S. 1023–1033.

Yona, Gal; Ghorbani, Amirata; Zou, James (2021): Who's Responsible? Jointly Quantifying the Contribution of the Learning Algorithm and Data. In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery, S. 1034–1041.

Zhang, Yunfeng; Bellamy, Rachel; Varshney, Kush (2020): Joint Optimization of AI Fairness and Utility: A Human-Centered Approach. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery, S. 400–406.

Zhang, Baobao; Dafoe, Allan (2020): U.S. Public Opinion on the Governance of Artificial Intelligence. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery, S. 187–193.

Zhang, Yue; Defazio, David; Ramesh, Arti (2021): RelEx: A Model-Agnostic Relational Model Explainer. In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery, S. 1042–1049.

Zhang, Chunkai; Du, Zilin; Yang, Yuting; Gan, Wensheng; Yu, Philip S. (2021): On-Shelf Utility Mining of Sequence Data. In: ACM Trans. Knowl. Discov. Data 16 (2). DOI: 10.1145/3457570.

Zhang, Chen; Hoi, Steven C. H.; Tsung, Fugee (2020): Time-Warped Sparse Non-Negative Factorization for Functional Data Analysis. In: ACM Trans. Knowl. Discov. Data 14 (6). DOI: 10.1145/3408313.

Zhang, Min-Ling; Wu, Jing-Han; Bao, Wei-Xuan (2022): Disambiguation Enabled Linear Discriminant Analysis for Partial Label Dimensionality Reduction. In: ACM Trans. Knowl. Discov. Data 16 (4). DOI: 10.1145/3494565.

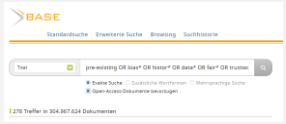
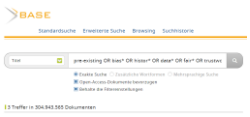
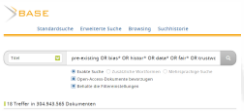
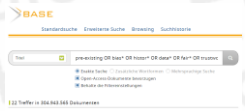
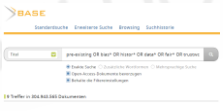
Zhou, Tongyu; Sheng, Haoyu; Howley, Iris (2020): Assessing Post-Hoc Explainability of the BKT Algorithm. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery, S. 407–413.

Zucker, Julian; d'Leeuwen, Myraeka (2020): Arbiter: A Domain-Specific Language for Ethical Machine Learning. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery, S. 421–425.

Zwetsloot, Remco; Zhang, Baobao; Dreksler, Noemi; Kahn, Lauren; Anderljung, Markus; Dafoe, Allan; Horowitz, Michael C. (2021): Skilled and Mobile: Survey Evidence of AI Researchers' Immigration Preferences. In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery, S. 1050–1059.

Anhang 2: BASE (Bielefeld Academic Search Engine)

Tab. II: Selektionsvorgehensweise der Publikationen Base

 *Trefferquote von 278 ohne Filter*		
Filter für Erscheinungsjahr	**Filter für Dokumentenart Artikel**	**Filter für Sprache Englisch**
 2022: 13 Treffer	 Artikel: 3 Treffer	 Sprache: 2
 2021: 42 Treffer	 Artikel: 22 Treffer	 Sprache: 17 Treffer
 2020: 35 Treffer	 Artikel: 18 Treffer	 Sprache: 12 Treffer
 2019: 22 Treffer	 Artikel: 11 Treffer	 Sprache: 9 Treffer
Gesamt		
112 Treffer	54 Treffer	40 Treffer

Literaturangaben zu Suchergebnissen in BASE (Bielefeld Academic Search Engine)

Allen, Robin; Masters, Dee (2019): Artificial Intelligence: the right to protection from discrimination caused by algorithms, machine learning and automated decision-making. DOI: 10.1007/s12027-019-00582-w.

Andreas Nieder (2020-08-01T00:00:00Z): Absolute Numerosity Discrimination as a Case Study in Comparative Vertebrate Intelligence. DOI: 10.3389/fpsyg.2020.01843.

Carter, Jeremy G.; Fox, Bryanna (2019): Community policing and intelligence-led policing; An examination of convergent or discriminant validity. DOI: 10.1108/pijpsm-07-2018-0105.

Carter, Jeremy G.; Fox, Bryanna (2019): Community Policing and Intelligence-Led Policing: An Examination of Convergent or Discriminant Validity. Online verfügbar unter http://hdl.handle.net/1805/20621.

Chen Li-Bin; Xu Dong-Mei; Chang Shih-Feng (2021-01-01T00:00:00Z): Using Artificial Intelligence to Explore Employment Discrimination and Interview Decision. DOI: 10.1051/e3sconf/202125101064.

Chinnasamy, Kavipriya; Arumugam, Yuvaraja; Jegadeesan, Ramalingam; Chockalingam, Vanniarajan (2020): Linear discriminant analysis in red sorghum using artificial intelligence. DOI: 10.1007/s13237-020-00340-1.

Dijun Rao; Xiuzhi Shi; Jian Zhou; Zhi Yu; Yonggang Gou; Zezhen Dong; Jinzhong Zhang (2021-07-01T00:00:00Z): An Expert Artificial Intelligence Model for Discriminating Microseismic Events and Mine Blasts. DOI: 10.3390/app11146474.

Eleonora Farina; Alessandro Pepe; Veronica Ornaghi; Valeria Cavioni (2021-07-01T00:00:00Z): Trait Emotional Intelligence and School Burnout Discriminate Between High and Low Alexithymic Profiles: A Study with Female Adolescents. DOI: 10.3389/fpsyg.2021.645215.

Farina, E.; Pepe, A.; Ornaghi, V.; Cavioni, V. (2021): Trait Emotional Intelligence and School Burnout Discriminate Between High and Low Alexithymic Profiles: A Study with Female Adolescents. Online verfügbar unter http://hdl.handle.net/10281/327218.

Gerards, Janneke; Zuiderveen Borgesius, Frederik (2020): Protected Grounds and the System of Non-Discrimination Law in the Context of Algorithmic Decision-Making and Artificial Intelligence. DOI: 10.2139/ssrn.3723873.

Gonçalves, Paulo J.S.; Estevinho, Letícia M.; Pereira, Ana Paula; Sousa, João M.C.; Anjos, O. (2019-04-12T23:28:49Z): Computational intelligence applied to discriminate bee pollen quality and botanical origin. Online verfügbar unter http://hdl.handle.net/10400.11/6458.

Heinrichs, Bert (2021): Discrimination in the age of artificial intelligence. DOI: 10.1007/s00146-021-01192-2.

Heinrichs, Bert (2022): Discrimination in the age of artificial intelligence. Online verfügbar unter https://juser.fz-juelich.de/record/891610.

Huanchi Wang; He Fang; Xianbin Wang (2021-01-01T00:00:00Z): Safeguarding Cluster Heads in UAV Swarm Using Edge Intelligence: Linear Discriminant Analysis-Based Cross-Layer Authentication. DOI: 10.1109/OJCOMS.2021.3084532.

Iman Mohammed; Shatha Mohammed (2019-12-01T00:00:00Z): A comparative Study for Designing an Efficient Intelligence System for the Process of Discrimination. DOI: 10.33899/edusj.1970.163336.

Jantzen, R.; Lin, F.; Abassade, P.; Billuart, O.; Antakly, Y.; Aroulanda, M. J. et al. (2020): Discriminating value of artificial intelligence based models for heart failure readmissions and mortality: A comparison of patients included or not in the PRADO program. DOI: 10.1016/j.acvdsp.2019.09.063.

Jastrzębski, Jan; Kroczek, Bartłomiej; Chuderski, Adam (2021): Galton and Spearman revisited: Can single general discrimination ability drive performance on diverse sensorimotor tasks and explain intelligence? DOI: 10.1037/xge0001005.

Kubišová, Milena; Pata, Vladimír; Měřínská, Dagmar; Škrobák, Adam; Marcaník, Miroslav (2021): Solving the issue of discriminant roughness of heterogeneous surfaces using elements of artificial intelligence. Online verfügbar unter http://publikace.k.utb.cz/handle/10563/1010366.

Kumbhar, D.; Palliyarayil, A.; Reghu, D.; Shrungar, D.; Umapathy, S.; Sil, S. (2021): Rapid discrimination of porous bio-carbon derived from nitrogen rich biomass using Raman spectroscopy and artificial intelligence methods. DOI: 10.1016/j.carbon.2021.03.064.

Laverde-Saad, Alexandra; Jfri, Abdulhadi; García, Rubén; Salgüero, Irene; Martínez, Constanza; Cembrero, Hirune et al. (2021): Discriminative deep learning based benignity/malignancy diagnosis of dermatologic ultrasound skin lesions with pretrained artificial intelligence architecture. DOI: 10.1111/srt.13086.

Loza de Siles, Emile (2020): AI, on Algorithmic Justice: A New Proposal Toward the Identification and Reduction of Discriminatory Bias in Artificial Intelligence Systems [Abstract]. DOI: 10.2139/ssrn.3658682.

Marco Peruzzi (2021-06-01T00:00:00Z): Anti-discrimination law and the challenge of artificial intelligence. DOI: 10.6092/issn.2421-2695/13117.

Milena Kubišová; Vladimír Pata; Dagmar Měřínská; Adam Škrobák; Miroslav Marcaník (2021-05-01T00:00:00Z): Solving the Issue of Discriminant Roughness of Heterogeneous Surfaces Using Elements of Artificial Intelligence. DOI: 10.3390/ma14102620.

Miller, S.R.M. (2021): Rethinking the Just Intelligence Theory of National Security Intelligence Collection and Analysis: The Principles of Discrimination, Necessity, Proportionality and Reciprocity. Online verfügbar unter http://resolver.tudelft.nl/uuid:133c09bd-04eb-409c-9012-cb11aa01d115.

Miller, Seumas (2021): Rethinking the Just Intelligence Theory of National Security Intelligence Collection and Analysis: The Principles of Discrimination, Necessity, Proportionality and Reciprocity. DOI: 10.1080/02691728.2020.1855484.

Moss, Haley (2021): Screened Out Onscreen: Disability Discrimination, Hiring Bias, and Artificial Intelligence. DOI: 10.2139/ssrn.3906300.

Nieder, Andreas (2020): Absolute Numerosity Discrimination as a Case Study in Comparative Vertebrate Intelligence. Online verfügbar unter http://hdl.handle.net/10900/109837.

Pelău, Corina; Ene, Irina (2020): Interaction Between Consumers and Emerging forms of Artificial Intelligence: A Discriminant Analysis. DOI: 10.2478/sues-2020-0008.

Pelău Corina; Ene Irina (2020-06-01T00:00:00Z): Interaction Between Consumers and Emerging forms of Artificial Intelligence: A Discriminant Analysis. DOI: 10.2478/sues-2020-0008.

Robinson, Michael D.; Persich, Michelle R.; Stawicki, Cassandra; Krishnakumar, Sukumarakurup (2019): Deviant Workplace Behavior as Emotional Action: Discriminant and Interactive Roles for Work-Related Emotional Intelligence. DOI: 10.1080/08959285.2019.1664548.

Sahar Mansour; Rasha Kamel; Ahmed Marey; Christiane Hunold; Ahmed Yousry (2022-03-01T00:00:00Z): Discrimination between phyllodes tumor and fibro-adenoma: Does artificial intelligence-aided mammograms have an impact? DOI: 10.1186/s43055-022-00734-y.

Saslow, Kate; Lorenz, Philippe (2019): Artificial Intelligence Needs Human Rights: How the Focus on Ethical AI Fails to Address Privacy, Discrimination and Other Concerns. DOI: 10.2139/ssrn.3589473.

Shestakova, Veronika (2021): Best Practices to Mitigate Bias and Discrimination in Artificial Intelligence. DOI: 10.1002/pfi.21987.

Todolí-Signes, Adrián (2019): Algorithms, artificial intelligence and automated decisions concerning workers and the risks of discrimination: the necessary collective governance of data protection. DOI: 10.1177/1024258919876416.

Tomasz Krzeszowski; Krzysztof Wiktorowicz (2020-11-01T00:00:00Z): Combined Regularized Discriminant Analysis and Swarm Intelligence Techniques for Gait Recognition. DOI: 10.3390/s20236794.

Tsukahara, Jason S.; Harrison, Tyler L.; Draheim, Christopher; Martin, Jessie D.; Engle, Randall W. (2020): Attention control: The missing link between sensory discrimination and intelligence. DOI: 10.3758/s13414-020-02044-9.

Tyagi, Ashish; Tiwari, Parul; Bhardwaj, Piyush; Chawla, Hitesh (2021): Prognosis of sexual dimorphism with unfused hyoid bone: Artificial intelligence informed decision making with discriminant analysis. DOI: 10.1016/j.scijus.2021.10.002.

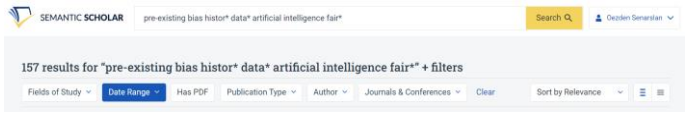
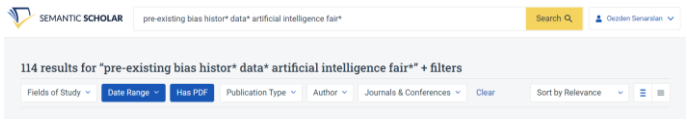
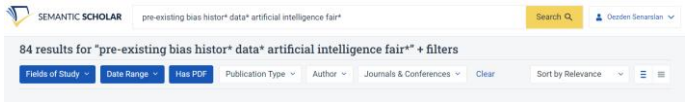
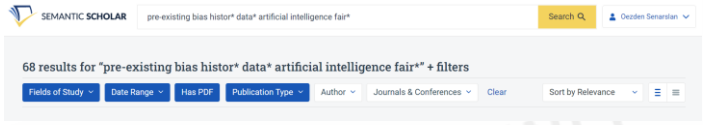
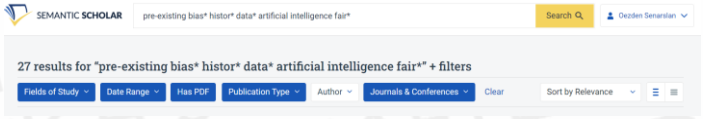
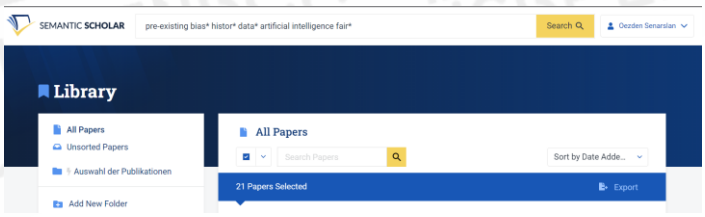
Wang, Shutao; Zhang, Demei (2020): The impact of perceived social support on students' pathological internet use: The mediating effect of perceived personal discrimination and moderating effect of emotional intelligence. DOI: 10.1016/j.chb.2020.106247.

Yavuz, Can (2019): Machine Bias: Artificial Intelligence and Discrimination. DOI: 10.2139/ssrn.3439702.

Zuiderveen Borgesius, Frederik J. (2020): Strengthening legal protection against discrimination by algorithms and artificial intelligence. DOI: 10.1080/13642987.2020.1743976.

Anhang 3: Semantic Scholar

Tab. III: Selektionsvorgehensweise der Publikationen Semantic Scholar

Filter	Suche	Treffer
Zeit: 2019-2022		157
Dateiformat: PDF		114
Field of Study: Computer Science, Philosophy, Sociology		84
Publication Type:		68
Journals & Conferences: ArXiv (21), AIES (4), KDD (2)		27
Nach Entfernung der Duplikate		21

Literaturangaben zu Suchergebnissen in Semantic Scholar

Akshat Pandey; Aylin Caliskan (2020): Iterative Effect-Size Bias in Ridehailing: Measuring Social Bias in Dynamic Pricing of 100 Million Rides. In: ArXiv abs/2006.04599.

Akshat Pandey; Aylin Caliskan (2021): Disparate Impact of Artificial Intelligence Bias in Ridehailing Economy's Price Discrimination Algorithms. In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society.

Bo Li; Peng Qi; Bo Liu; Shuai Di; Jingen Liu; Jiquan Pei et al. (2021): Trustworthy AI: From Principles to Practices. In: ArXiv abs/2110.01167.

Caitlin Kuhlman; Latifa Jackson; Rumi Chunara (2020): No Computation without Representation: Avoiding Data and Algorithm Biases through Diversity. In: Proceedings

of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining.

Deepa Muralidhar (2021): Examining Religion Bias in AI Text Generators. In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society.

Kacper Sokol (2019): Fairness, Accountability and Transparency in Artificial Intelligence: A Case Study of Logical Predictive Models. In: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society.

Krzysztof Chomiak; Michal Miktus (2021): Harnessing value from data science in business: ensuring explainability and fairness of solutions. In: ArXiv abs/2108.07714.

Mariya I. Vasileva (2020): The Dark Side of Machine Learning Algorithms: How and Why They Can Leverage Bias, and What Can Be Done to Pursue Algorithmic Fairness. In: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining.

Mehdi Yazdani-Jahromi; Amirarsalan Rajabi; Aida Tayebi; Ozlem Ozmen Garibay (2022): Distraction is All You Need for Fairness. In: ArXiv abs/2203.07593.

Melanie Bernhardt; Charles Jones; Ben Glocker (2022): Investigating underdiagnosis of AI algorithms in the presence of multiple sources of dataset bias. In: ArXiv abs/2201.07856.

Nicholas Schmidt; Bryce E. Stephens (2019): An Introduction to Artificial Intelligence and Solutions to the Problems of Algorithmic Discrimination. In: ArXiv abs/1911.05755.

Ninareh Mehrabi; Umang Gupta; Fred Morstatter; Greg Ver Steeg; A. G. Galstyan (2021): Attributing Fair Decisions with Attention Interventions. In: ArXiv abs/2109.03952.

Nuri Mahmoud Ahmed; Muntasir Wahed (2020): The De-democratization of AI: Deep Learning and the Compute Divide in Artificial Intelligence Research. In: ArXiv abs/2010.15581.

Romila Pradhan; Jiongli Zhu; Boris Glavic; Babak Salimi (2022): Interpretable Data-Based Explanations for Fairness Debugging. In: Proceedings of the 2022 International Conference on Management of Data.

Shubham Sharma; Alan H. Gee; David Paydarfar; Joydeep Ghosh (2021): FaiR-N: Fair and Robust Neural Networks for Structured Data. In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society.

Susan Leavy; Barry O'Sullivan; Eugenia Siapera (2020): Data, Power and Bias in Artificial Intelligence. In: ArXiv abs/2008.07341.

Tal Feldman; Ashley Peake (2021): On the Basis of Sex: A Review of Gender Bias in Machine Learning Applications. In: ArXiv abs/2104.02532.

Tatiana Tommasi; Silvia Bucci; Barbara Caputo; Pietro Asinari (2021): Towards Fairness Certification in Artificial Intelligence. In: ArXiv abs/2106.02498.

Xiaoxiao Li; Ziteng Cui; Yifan Wu; Li Gu; Tatsuya Harada (2021): Estimating and Improving Fairness with Adversarial Learning. In: ArXiv abs/2103.04243.

Yukun Zhang; Longsheng Zhou (2019): Fairness Assessment for Artificial Intelligence in Financial Industry. In: ArXiv abs/1912.07211.

Zhibo Wang; Xiaowei Dong; Henry Xue; Zhifei Zhang; Weifeng Chiu; Tao Wei; Kui Ren (2022): Fairness-aware Adversarial Perturbation Towards Bias Mitigation for Deployed Deep Models. In: ArXiv abs/2203.01584.

Anhang 4: Ausgeschlossene Publikationen nach Volltextanalyse

Tab. IV: Ausgeschlossene Publikationen nach Volltextanalyse

ID	Titel	Autor*innen	Kommentar für den Ausschluss	Jahr	Punkte
_P1	Persistent Anti-Muslim Bias in Large Language Models	Abid, Abubakar; Farooqi, Maheen; Zou, James	Technische Lösung zu Sprachdiskriminierung	2021	2
_P6	Iterative Effect-Size Bias in Ridehailing: Measuring Social Bias in Dynamic Pricing of 100 Million Rides	Akshat Pandey; Aylin Caliskan	Technische Lösung zu Preisbestimmungen bei Personenbeförderung wie Uber.	2020	2
_P15	Exploring AI Futures Through Role Play	Avin, Shahar; Gruetzemacher, Ross; Fox, James	Methodik, die Auswirkungen von KI anhand eines Rollenspiels untersucht und lehrt. Keine relevanten Ergebnisse für Daten. Es geht um Rollenspiele, die Machtpositionen wie CEO, Präsident*in der USA usw. bekleiden und Entscheidungen treffen sollen.	2020	1,5

ID	Titel	Autor*innen	Kommentar für den Ausschluss	Jahr	Punkte
_P19	TrolleyMod v1.0: An Open-Source Simulation and Data-Collection Platform for Ethical Decision Making in Autonomous Vehicles	Behzadan, Vahid; Minton, James; Munir, Arslan	TrolleyMod v1.0 Vorgestellung. Ist eine Open-Source-Plattform auf der Grundlage des CARLA-Simulators für die Erfassung ethischer Entscheidungsdaten für autonome Fahrzeuge.Hat nichts mit historischen Daten zu tun.	2019	1,5
_P20	Activism by the AI Community: Analysing Recent Achievements and Future Prospects	Belfield, Haydn	Überblick über den Aktivismus der KI-Gemeinschaft in den letzten sechs Jahren. Laut Publikation hängt der bisherige Erfolg von einer kohärenten gemeinsamen Kultur und einer hohen Verhandlungs-macht aufgrund der hohen Nachfrage nach einem begrenzten Angebot an KI-"Talenten" ab. Keine relevanten Info zu Daten!	2020	1,5

ID	Titel	Autor*innen	Kommentar für den Ausschluss	Jahr	Punkte
_P21	Automating Procedurally Fair Feature Selection in Machine Learning	Belitz, Clara; Jiang, Lan; Bosch, Nigel	Methode zur Auswahl von Merkmalen, die sich nicht nur an fairen Ergebnissen, sondern auch an fairen Prozessen (prozedurale Fairness) orientiert. Mathematische Lösungsvorschlag, wie Fainess in Daten über Metriken eingestellt werden.	2021	1,5
_P22	Explainable AI and Adoption of Financial Algorithmic Advisors: An Experimental Study	Ben David, Daniel; Resheff, Yehezkel S.; Tron, Talia	Untersuchung der Akzeptanz von menschlichen und KI-Finanzberater*in durch ein webbasiertes Spiel mit realen monetären Konsequenzen. Der Großteil dieser Arbeit konzentriert sich auf algorithmische Methoden zur Erstellung von Erklärungen für komplexe maschinelle Lernmethoden.	2021	1,5
_P33	The Problem with Intelligence: Its Value-Laden History and the Future of AI	Cave, Stephen	In diesem Beitrag wird argumentiert, dass das Konzept der Intelligenz in einer Weise wertgeladen ist, die sich auf den Bereich der KI und die Debatten über ihre Risiken und Chancen auswirkt.	2020	2

ID	Titel	Autor*innen	Kommentar für den Ausschluss	Jahr	Punkte
_P34	Scary Robots": Examining Public Responses to AI	Cave, Stephen; Coughlan, Kate; Dihal, Kanta	Wahrnehmung von KI in der britischen Öffentlichkeit. Hat aber nichts mit Daten zu tun. In diesem Beitrag wurde eine landesweit repräsentative Umfrage unter der britischen Bevölkerung zu ihrer Wahrnehmung von KI untersucht, wobei der Schwerpunkt auf den Gefühlen gegenüber utopischen oder dystopischen Zukunfts-szenarien lag. Insgesamt zeigen die Ergebnisse, dass die Bevölkerung eine ausgesprochen negative Einstellung zu dieser Technologie hat: Der Grad der Besorgnis war bei allen Erzählungen im Durchschnitt höher als der Grad der Begeisterung;	2019	1,5
_P39	A Just Approach Balancing Rawlsian Leximax Fairness and Utilitarianism	Chen, Violet; Hooker, J. N.	Modellierung und Definition von Fairness-Effizienz-Ausgleich, um Ressourcenzuteilung mit Fairness und Effizienz auszugleichen. Motiviert durch das gerechtigkeitsorientierte Abwägungsprinzip zwischen Fairness und Effizienz, d.h. man sollte der fairen Behandlung von Benachteiligten so lange Vorrang einräumen, bis ein zu großer Verzicht auf Effizienz erforderlich ist, entwickeln wir einen sequenziellen Optimierungsansatz, um Rawls'sche Leximax-Fairness und Utilitarismus auszugleichen.	2020	1,5

ID	Titel	Autor*innen	Kommentar für den Ausschluss	Jahr	Punkte
_P40	Gender Bias and Under-Representation in Natural Language Processing Across Human Languages	Chen, Yan; Mahoney, Christopher; Grasso, Isabella; Wali, Esma; Matthews, Abigail; Middleton, Thomas; Njie, Mariama; Matthews, Jeanna	Verarbeitung von natürlicher Sprache (Natural Language Processing, NLP) durch KI-Anwendungen. Untersuchung von neuen Sprachen auf geschlechtsspezifische Verzerrungen in den Wikipedia-Foren. Hat leider nichts zu Daten stehen, die für die vorliegende Arbeit relevant wären.	2021	1,5
_P44	Should Artificial Intelligence Governance Be Centralised? Design Lessons from History	Cihon, Peter; Maas, Matthijs M.; Kemp, Luke	Die Publikation geht der Frage nach, ob eine wirksame internationale Governance für KI-Anwendungen fragmentiert bleiben kann, oder ist eine zentralisierte internationale Organisation für KI-Anwendungen erforderlich. Keine relevanten Infos zu Daten.	2020	1
_P60	Algorithmic Fairness from a Non-Ideal Perspective	Fazelpour, Sina; Lipton, Zachary C.	In dieser Publikation wird eine Verbindung zwischen dem Ansatz des fairen maschinellen Lernverfahrens und der Literatur über ideale und nicht-ideale methodologische Ansätze in der politischen Philosophie hergestellt. Es geht bei dieser Publikation mehrheitlich um Algorithmen.	2020	1,5

ID	Titel	Autor*innen	Kommentar für den Ausschluss	Jahr	Punkte
_P62	Reflexive Design for Fairness and Other Human Values in Formal Models	Fish, Benjamin; Stark, Luke	Aufzeigen von konzeptionellen Grenzen der formalen Modellierung einer KI-Anwendung und Entwicklung von vier reflexiven Werten - Wertetreue, angemessene Genauigkeit, Lesbarkeit von Werten und Anfechtung von Werten die für eine angemessene Einbeziehung menschlicher Werte in formale Modelle unerlässlich sind. Interessante Publikation, wenn um reflexives Designing von Entwickelnden geht. Hat aber nichts mit Daten zu tun.	2021	1,5
_P69	Protected Grounds and the System of Non-Discrimination Law in the Context of Algorithmic Decision-Making and Artificial Intelligence	Gerards, Janneke; Zuiderveen Borgesius, Frederik	In diesem Papier wird untersucht, welches System des Antidiskriminierungsrechts am besten auf die algorithmische Entscheidungsfindung von KI-Anwendungen angewandt werden können, weil Algorithmen auf der Grundlage von Merkmalen differenzieren können, die nicht mit geschützten Diskriminierungsmerkmalen wie ethnischer Herkunft oder Geschlecht korrelieren. In dieser Publikation geht es um die Antidiskriminierungsgesetze. Es geht nicht um Daten.	2020	1,5

ID	Titel	Autor*innen	Kommentar für den Ausschluss	Jahr	Punkte
_P73	Quality-Informed Process Mining: A Case for Standardised Data Quality Annotations	Goel, Kanika; Leemans, Sander J. J.; Martin, Niels; Wynn, Moe T.	Es wird eine Datenqualitätsaussage für Ereignisprotokolle vorgeschlagen, die von Process-Mining-Algorithmen verwendet werden können, um qualitätsbezogene Erkenntnisse zu gewinnen. Mithilfe eines Design-Science-Ansatzes werden Anforderungen formuliert, die für den Vorschlag von Datenqualitäts-Aussagen genutzt werden. Es geht um Process Mining die strukturierte Erfassung von Metadaten zur Datenqualität, die von Algorithmen genutzt werden können.	2022	2
_P88	Discrimination in the age of artificial intelligence	Heinrichs, Bert	Die Publikation definiert zuerst selbst, was Diskriminierung ist und behauptet, dass ADM's in KI-Anwendungen Diskriminierung nicht verschärfen, jedoch tatsächlich dabei helfen können, versteckte Formen der Diskriminierung aufzudecken. Es geht mehr um ADM's als um Daten.	2022	1,5

ID	Titel	Autor*innen	Kommentar für den Ausschluss	Jahr	Punkte
_P89	Situated Accountability: Ethical Principles, Certification Standards, and Explanation Methods in Applied AI	Henriksen, Anne; Enni, Simon; Bechmann, Anja	In dieser Publikation werden drei Mechanismen untersucht, indem sie den Umgang skandinavischer KI-Entwickelnden mit (1) ethischen Prinzipien, (2) Zertifizierungsstandards und (3) Erklärungsmethoden analysiert. Die Publikation thematisiert nicht Daten, sondern macht ein ethnographic case study conducted at an AI company in Scandinavia.	2021	1,5
_P91	Measuring Fairness in an Unfair World	Herington, Jonathan	Diese Publikation schlägt einen alternativen Rahmen zur Messung von Fairness im Kontext bestehender Ungerechtigkeit vor, die Verteilungsgerechtigkeit beinhalten soll. Die Publikation will messen.	2020	1,0
_P107	Ethics of Food Recommender Applications	Karpati, Daniel; Najjar, Amro; Ambrossio, Diego Agustin	Die Publikation setzt sich mit Food Recommender Systems auseinander und hat die größten Herausforderungen identifiziert und schlägt ein Schema vor, wie explizite ethische Agenden erklärt werden sollten.	2020	1,5

ID	Titel	Autor*innen	Kommentar für den Ausschluss	Jahr	Punkte
_P121	A Framework for Benchmarking Discrimination-Aware Models in Machine Learning	L. Cardoso, Rodrigo; Meira Jr., Wagner; Almeida, Virgilio; J. Zaki, Mohammed	Die Publikation schlägt ein Benchmark-Rahmenwerk zur Bewertung von diskriminierungs-sensitiven Modellen vor. Der Rahmen besteht aus systematisch generierten, verzerrten Daten-sätzen, die den realen Daten ähneln und mit einem Bayes'schen Netzwerkansatz erstellt werden. Das Rahemnwerk kann wohl auf die meisten Datensätze angewandt werden.	2019	1,5
_P128	Normative Principles for Evaluating Fairness in Machine Learning	Leben, Derek	Ziel dieses Beitrags ist, die Erfolgs- und Fehlerquoten für geschützte Gruppen (Race, Gender, sexuelle Orientierung) als ein Verteilungs-problem zu charakterisieren und die möglichen Lösungen für dieses Problem anhand verschiedener normativer Prinzipien aus der moralischen und politischen Philosophie zu beschreiben.	2020	1,5

ID	Titel	Autor*innen	Kommentar für den Ausschluss	Jahr	Punkte
_P143	Ethics for AI Writing: The Importance of Rhetorical Context	McKee, Heidi A.; Porter, James E.	Diese Publikation schlägt zwei ethische Prinzipien vor, die das Design von KI-Anwendungen leiten sollen: Transparenz über die maschinelle Präsenz und kritisches Datenbewusstsein, eine methodologische Reflexivität über rhetorischen Kontext und Auslassungen in den Daten, die von einem menschlichen Agenten bereitgestellt oder beim maschinellen Lernen berücksichtigt werden müssen. Es geht hierbei um Text-basierte AI-Systeme	2020	1,5
_P144	Costs and Benefits of Fair Representation Learning	McNamara, Daniel; Ong, Cheng Soon; Williamson, Robert C.	Die Publikation zählt die Vorteile des maschinellen Lernens durch einer fairen Datenrepräsentation auf und zeigt auf, dass jede nachfolgende Nutzung der bereinigten Daten nicht unfair ist.	2019	1,5
_P146	Investigating underdiagnosis of AI algorithms in the presence of multiple sources of dataset bias	Melanie Bernhardt; Charles Jones; Ben Glocker	In dieser Publikation geht es um Brustkorb-Röntgendatensätzen, mit denen KI-Anwendungen trainiert werden und falsche Diagnosen aufgrund von Gruppen-Unterrepräsention treffen. Hier wird die Diagnose der KI-Anwendungen diskutiert.	2022	1,0

ID	Titel	Autor*innen	Kommentar für den Ausschluss	Jahr	Punkte
_P147	Causal Multi-Level Fairness	Mhasawade, Vishwali; Chunara, Rumi	Diese Publikationen thematisiert auf Makroebene die strukturelle Diskriminierung und ihre Auswirkung durch KI-Anwendungen und stellet einen Ansatz zur Vorhersage von Einkommen auf der Grundlage von Attributen auf Makro- und Individualebene vor, die die Unfairness abschwächen sollen.	2021	1,5
_P155	An Introduction to Artificial Intelligence and Solutions to the Problems of Algorithmic Discrimination	Nicholas Schmidt; Bryce E. Stephens	Diese Publikation gibt einen Überblick über die potenziellen Vorteile und Risiken, die mit der Verwendung von Algorithmen und Daten verbunden sind, und konzentriert dabei speziell auf die Fairness von Konsumentenkrediten und auf die rechtlichen Anforderungen des Equal Credit and Opportunity Act.	2019	1,5
_P158	Attributing Fair Decisions with Attention Interventions	Ninareh Mehrabi; Umang Gupta; Fred Morstatter; Greg Ver Steeg; A. G. Galstyan	In dieser Publikation wird ein Ansatz vorgestellt, der mit zwei verschiedenen Datentypen, tabellarischen und textuellen Daten experimentiert. Das Ziel ist ein aufmerksamkeitsbasiertes Modell zu entwerfen, das als Framework für die Attribution genutzt werden kann.	2021	1,5

ID	Titel	Autor*innen	Kommentar für den Ausschluss	Jahr	Punkte
_P159	Active Fairness in Algorithmic Decision Making	Noriega-Campero, Alejandro; Bakker, Michiel A.; Garcia-Bulle, Bernardo; Pentland, Alex 'Sandy'	Die Publikation schlägt einen alternativen aktiven Rahmen für eine faire Klassifizierung vor, bei dem eine Entscheidungsträger*in im Einsatz adaptiv Informationen entsprechend den Bedürfnissen verschiedener Gruppen oder Einzelpersonen erwirbt, um Ungleichheiten in der Klassifizierungsleistung auszugleichen. Die Publikation thematisiert zwei solcher Methoden, bei denen die Informationsbeschaffung an die Bedürfnisse der Gruppe bzw. des Einzelnen angepasst wird.	2019	1,5
_P174	Actionable Auditing: Investigating the Impact of Publicly Naming Biased Performance Results of Commercial AI Products	Raji, Inioluwa Deborah; Buolamwini, Joy	Diese Publikation analysiert die Auswirkungen der öffentlichen Benennung und Offenlegung von Leistungsergebnissen voreingenommener KI-Systeme und untersucht die kommerziellen Auswirkungen von Gender Shades, dem ersten Algorithmus-Audit von geschlechts- und hauttypbezogenen Leistungsunterschieden in kommerziellen Gesichtsanalysemodellen.	2019	1,5

ID	Titel	Autor*innen	Kommentar für den Ausschluss	Jahr	Punkte
_P175	Saving Face: Investigating the Ethical Concerns of Facial Recognition Auditing	Raji, Inioluwa Deborah; Gebru, Timnit; Mitchell, Margaret; Buolamwini, Joy; Lee, Joonseok; Denton, Emily	Diese Publikation zeigt fünf ethische Bedenken hinsichtlich der Auditerung von kommerziellen Gesichtserkennungstechnologien auf und hebt zusätzliche Designüberlegungen und ethische Spannungen hervor, denen sich die Prüfenden bewusst sein sollen, um die von dem geprüften System verbreiteten Schäden nicht zu verschlimmern oder zu ergänzen.	2020	1,5
_P181	Artificial Intelligence Needs Human Rights: How the Focus on Ethical AI Fails to Address Privacy, Discrimination and Other Concerns	Saslow, Kate; Lorenz, Philippe	Diese Publikation plädiert für einen Perspektivenwechsel. Anstatt ethische Werte für KI-Anwendungen zu definieren, sollten demnach KI-Anwendungen aus einer Menschenrechtsperspektive untersucht werden, um aktuelle und künftige Schäden zu analysieren, die durch KI verursacht oder verschlimmert werden, und um Maßnahmen zur Vermeidung von Schäden zu ergreifen. Daher sollen Unternehmen und Staaten sich um die Entwicklung von KI-Technologien bemühen, die die Menschenrechte wahren.	2019	1,5

ID	Titel	Autor*innen	Kommentar für den Ausschluss	Jahr	Punkte
_P188	Fairness in the Eyes of the Data: Certifying Machine-Learning Models	Segal, Shahar; Adi, Yossi; Pinkas, Benny; Baum, Carsten; Ganesh, Chaya; Keshet, Joseph	Die Publikation stellt einen Rahmen vor, der es ermöglicht, den Fairnessgrad eines Modells auf der Grundlage eines interaktiven und die Privatsphäre wahrenden Tests zu zertifizieren. Der Rahmen verifiziert jedes trainierte Modell, unabhängig von seinem Trainingsprozess und seiner Architektur.	2021	1,5
_P193	FairOD: Fairness-Aware Outlier Detection	Shekhar, Shubhranshu; Shah, Neil; Akoglu, Leman	Diese Publikation hat einen fairnessbewussten Ausreißerdetektor (Fairness Outlier Detection (FairOD)), der die folgenden Eigenschaften aufweist: FairOD (1) weist Behandlungsparität zum Testzeitpunkt auf, (2) zielt darauf ab, gleiche Anteile von Proben aus allen Gruppen zu markieren (d.h. Gruppenfairness durch statistische Parität zu erreichen), und (3) strebt danach, wirklich risikoreiche Proben innerhalb jeder Gruppe zu markieren.	2021	1,5

ID	Titel	Autor*innen	Kommentar für den Ausschluss	Jahr	Punkte
_P208	On the Basis of Sex: A Review of Gender Bias in Machine Learning Applications	Tal Feldman; Ashley Peake	Die Publikation gibt eine Übersicht, die sich speziell mit geschlechtsspezifischen Verzerrungen in KI-Anwendungen des maschinellen Lernens befasst. Es werden einige Beispiele für geschlechtsspezifische Verzerrungen beim maschinellen Lernen in der Praxis vorgestellt. Außerdem werden mehrere Open-Source-Pakete für Fairness in der KI besprochen.	2021	
_P211	Algorithms, artificial intelligence and automated decisions concerning workers and the risks of discrimination: the necessary collective governance of data protection	Todolí-Signes, Adrián	In dieser Publikation werden die in der EU-Datenschutz-Grundverordnung (DSGVO) festgelegten Schutzmaßnahmen zum Schutz der Arbeitnehmer*innen vor Diskriminierung analysiert.	2019	1,5
_P215	Social and Governance Implications of Improved Data Efficiency	Tucker, Aaron D.; Anderljung, Markus; Dafoe, Allan	Laut der vorliegenden Publikation arbeiten viele Forschende daran, die Dateneffizienz des maschinellen Lernens zu verbessern. Was würde passieren, wenn sie Erfolg hätten? In dieser Publikation werden die sozioökonomischen Auswirkungen einer erhöhten Dateneffizienz untersucht.	2020	1,5

ID	Titel	Autor*innen	Kommentar für den Ausschluss	Jahr	Punkte
_P227	Who’s Responsible? Jointly Quantifying the Contribution of the Learning Algorithm and Data	Yona, Gal; Ghorbani, Amirata; Zou, James	Diese Publikation beschäftigt sich mit der Frage, wo die Verantwortung bei Fehlentscheidungen liegen sollte, wenn ein Lernalgorithmus A, der auf einem Datensatz D trainiert wurde, sich zum Testzeitpunkt bei einer Teilpopulation als schlecht erweist.	2021	1,5

Tab. V: Eingeschlossene Publikationen nach Volltextanalyse

ID	Titel	Publizierende	Jahr	Punkte
P2	Are AI Ethics C onferences Different and More Diverse Compared to Traditional Computer Science Conferences?	Acuna, Daniel E.; Liang, Lizhen	2021	1,5
P10	AI and Holistic Review: Informing Human Reading in College Admissions	Alvero, A. J.; Arthurs, Noah; antonio, anthony lising; Domingue, Benjamin W.; Gebre-Medhin, Ben; Giebel, Sonia; Stevens, Mitchell L.	2020	1,5
P18	Designing Disaggregated Evaluations of AI Systems: Choices, Considerations, and Tradeoffs	Barocas, Solon; Guo, Anhong; Kamar, Ece; Krones, Jacquelyn; Morris, Meredith Ringel; Vaughan, Jennifer Wortman; Wadsworth, W. Duncan; Wallach, Hanna	2021	1,5
P23	Artificial Intelligence and the Purpose of Social Systems	Benthall, Sebastian; Goldenfein, Jake	2021	1,5
P28	Envisioning Communities: A Participatory Approach Towards AI for Social Good	Bondi, Elizabeth; Xu, Lily; Acosta-Navas, Diana; Killian, Jackson A.	2021	1,5
P29	A Comparative Analysis of Emotion-Detecting AI Systems with Respect to Algorithm Performance and Dataset Diversity	Bryant, De'Aira; Howard, Ayanna	2019	1,5
P31	No Computation without Representation: Avoiding Data and Algorithm Biases through Diversity	Caitlin Kuhlman; Latifa Jackson; Rumi Chunara	2020	2,0
P41	Reconfiguring Diversity and Inclusion for AI Ethics	Chi, Nicole; Lurie, Emma; Mulligan, Deirdre K.	2021	1,5
P45	Emergent Unfairness in Algorithmic Fairness-Accuracy Trade-Off Research	Cooper, A. Feder; Abrams, Ellen; NA, N. A.	2021	2,0

ID	Titel	Publizierende	Jahr	Punkte
P47	Theories of Parenting and Their Application to Artificial Intelligence	Croeser, Sky; Eckersley, Peter	2019	1,5
P49	An Invitation to System-Wide Algorithmic Fairness	Cruz Cortés, Efrén; Ghosh, Debashis	2020	1,5
P65	Auditing Algorithms: On Lessons Learned and the Risks of Data Minimization	Galdon Clavell, Gemma; Mart\'ın Zamorano, Mariano; Castillo, Carlos; Smith, Oliver; Matic, Aleksandar	2020	1,5
P71	Epistemic Therapy for Bias in Automated Decision-Making	Gilbert, Thomas Krendl; Mintz, Yonatan	2019	1,5
P106	The Global South is Everywhere, but Also Always Somewhere: National Policy Narratives and AI Justice	Kak, Amba	2020	1,5
P117	Becoming Good at AI for Good	Kshirsagar, Meghana; Robinson, Caleb; Yang, Siyu; Gholami, Shahrzad; Klyuzhin, Ivan; Mukherjee, Sumit; Nasir, Md; Ortiz, Anthony; Oviedo, Felipe; Tanner, Darren; Trivedi, Anusua; Xu, Yixi; Zhong, Ming; Dilkina, Bistra; Dodhia, Rahul; Lavista Ferres, Juan M.	2021	1,5
P127	Ethical Data Curation for AI: An Approach Based on Feminist Epistemology and Critical Theories of Race	Leavy, Susan; Siapera, Eugenia; O'Sullivan, Barry	2021	2,0
P129	Risk Identification Questionnaire for Detecting Unintended Bias in the Machine Learning Development Lifecycle	Lee, Michelle Seng Ah; Singh, Jatinder	2021	1,5
P140	Lessons from Archives: Strategies for Collecting Sociocultural Data in Machine Learning.	Eun Seo Jo, Timnit Gebru	2019	2,0
P165	Understanding the Representation and Representativeness of Age in AI Data Sets	Park, Joon Sung; Bernstein, Michael S.; Brewer, Robin N.; Kamar, Ece; Morris, Meredith Ringel	2021	1,5

ID	Titel	Publizierende	Jahr	Punkte
P170	We Haven’t Gone Paperless Yet: Why the Printing Press Can Help Us Understand Data and AI	Posada, Julian; Weller, Nicholas; Wong, Wendy H.	2021	2,0
P184	The Theory, Practice, and Ethical Challenges of Designing a Diversity-Aware Platform for Social Relations	Schelenz, Laura; Bison, Ivano; Busso, Matteo; Götzen, Amalia de; Gatica-Perez, Daniel; Giunchiglia, Fausto; Meegahapola, Lakmal; Ruiz-Correa, Salvador	2021	1,5
P192	Data Augmentation for Discrimination Prevention and Bias Disambiguation	Sharma, Shubham; Zhang, Yunfeng; R\’ıos Aliaga, Jesús M.; Bouneffouf, Djallel; Muthusamy, Vinod; Varshney, Kush R.	2020	2,0
P204	Data, Power and Bias in Artificial Intelligence	Susan Leavy; Barry O’Sullivan; Eugenia Siapera	2020	2,0
P209	Towards Fairness Certification in Artificial Intelligence	Tatiana Tommasi; Silvia Bucci; Barbara Caputo; Pietro Asinari	2021	1,5
P212	Fairness for Unobserved Characteristics: Insights from Technological Impacts on Queer Communities	Tomasev, Nenad; McKee, Kevin R.; Kay, Jackie; Mohamed, Shakir	2021	1,5
P219	Alienation in the AI-Driven Workplace	Vredenburgh, Kate	2021	1,5
P222	The Role and Limits of Principles in AI Ethics: Towards a Focus on Tensions	Whittlestone, Jess; Nyrup, Rune; Alexandrova, Anna; Cave, Stephen	2019	1,5
P238	Strengthening legal protection against discrimination by algorithms and artificial intelligence	Zuiderveen Borgesius, Frederik J.	2020	1,5

Anhang 6: Kurze Vorstellung der ausgewählten Publikationen

Tab. VI: Ziele & Ergebnisse der ausgewählten Publikationen

ID	Ziel	Ergebnis
P2	Acuna et al. (2021): Entwicklung von AI ethics‘ (AIE) Konferenzen auf demografische Merkmale, Publikationsinhalte und Zitationsmuster.	Vielfältige und differenzierte Themen vertreten, die andere Informatik Veranstaltungen beeinflussen. *Weiße* Forschende häufiger vertreten als ältere und Schwarze Forschende.
P10	Alvero et al. (2020): Untersuchung von 283.676 Bewerbungsaufsätze als Textmaterial, die im Rahmen von Hochschul-zulassungen zwischen 2015 und 2016 in den USA eingereicht wurden.	KI-Anwendungen sind in der Lage, aus dem Textmaterial Geschlecht und Haushaltseinkommen mit hoher Genauigkeit vorherzusagen.
P18	Barocas et al. (2019): Disaggregierte Evaluierung, bei denen die Systemleistung von KI-Anwendungen für verschiedene Personengruppen getrennt untersucht wird.	Sie plädieren für die Dokumentation der Evaluierung und zum anderen für die Veröffentlichung dieser Dokumentation, um anderen Beteiligten Hilfestellung zu geben.
P23	Benthall et al. (2021): Analyse von rechtlichen Regulierungslösungen und dem Liberalismus, die in Zusammenhang mit der Fairness in KI-Anwendungen stehen.	KI-Anwendungen können nur insoweit ethisch sein, wie der Zweck des sozialen Systems, das sie betreibt. Vorstellung eines Genossenschaftsmodells, das selbstverwaltend als soziales System über die eigenen Daten entscheidet.
P28	Bondi et al. (2021): Gestaltung und Umsetzungsmöglichkeiten von soziotechnischen KI-Anwendungen.	Es wird Participatory Design Ansatz empfohlen, wo Mitglieder einer Gemeinschaft als gleichwertige Partner*innen befähigt werden bei der KI-Entwicklung mitzuwirken.
P29	Bryant et al. (2019): untersuchen Emotionsklassifizierungssysteme mit mehreren Datensätzen, die emotionale Gesichtsausdrücke von Kindern beinhalten und vergleichen diese mit Datensätzen von Erwachsenen.	Die Gesichtsausdrücke von Kindern werden viel schlechter erkannt werden als von Erwachsenen und unterbreiten Vorschläge zur Verbesserung.
P31	Kuhlman et al. (2020): Untersuchen verschiedene Datensätze	Sie stellen einen Zusammenhang her zwischen der mangelnden Diversität in der akademischen und professionellen Informatik und der Art und dem Umfang der Verzerrungen, die in Datensätzen, Problemformulierungen und der Interpretation der Ergebnisse auftreten.

P41	Chi et al. (2021): Dokumenten-analyse von Big-Tech und verwenden zu Analysezwecken den Value Design Ansatz, um das Verständnis von Diversität und Inklusion von Google, Microsoft und Salesforce in ihren öffentlich zugänglichen KI-Ethik-dokumentation zu untersuchen.	Das Ergebnis der Dokumentenanalyse ist, dass die Dokumente die Themen Diversität und Inklusion zwar für Ingenieur*innen und technische Kund*innen verständlicher machen, jedoch eine Tendenz zur Abkehr vom Zivilrecht erkennen lassen.
P45	Cooper et al. (2021) argumentieren, dass Forschende eindeutige mathematische Annahmen aufstellen, jedoch für die Optimierung von Fairness und Genauigkeit keine ähnliche Aufmerksamkeit auf die normativen Annahmen richten.	Kommen zum Ergebnis, dass solche Annahmen, die oft implizit und ungeprüft bleiben, zu widersprüchlichen Schlussfolgerungen führen.
P47	Croeser et al. (2019) gehen davon aus, dass die Forschung irgendwann künstliche allgemeine Intelligenz (AGI) schaffen wird, die mit einer dem Menschen vergleichbaren Offenheit und Autonomie denkt und handelt.	Sie argumentieren, dass in die Forschungsbemühungen der Blickwinkel, um die der radikalen Elternschaft zu erweitern.
P49	Cruz Cortés et al. (2020) schlagen agentenbasierte Modelle als Erklärungsinstrumente für algorithmische Fairness vor.	Der Kerngedanke ist, dass für jede Gruppe in einem Datensatz ein Agent zugewiesen wird, um durch Simulation die Interaktion und das Verhalten einer KI-Anwendung zu verstehen
P65	Galdon Clavell et al. (2020) stellen die App Algorithmen-Audit (AA) von REM!X vor.	Das Hauptziel der App ist, algorithmische Verzerrungen im Empfehlungssystem zu identifizieren und abzuschwächen, die zur Diskriminierung geschützter Gruppen führen könnten. Die Publizierenden zeigen die Operationalisierung solcher Apps auf sowie deren Potenzial und ihre Grenzen.
P71	Gilbert et al. (2019) führen das Konzept ‚alief-discordant belief diagnoses' ein.	Ihre Erörterung diagnostiziert die ethischen Probleme, die bei der Entwicklung von KI-Anwendungen auf der Grundlage menschlicher Voreingenommenheit entstehen.
P106	Kak (2020) thematisiert die Notwendigkeit die Aufmerksamkeit auf den Globalen Süden zu lenken, da der Mainstream-Diskurs des Globalen Nordens über KI-Ethik parallel zu den Forderungen der Regierungen und Gemeinschaften verläuft.	Die Publikation unterstreicht die Notwendigkeit, die politische Ökonomie der KI aus verschiedenen Blickwinkeln zu betrachten - auf globaler und nationaler Ebene sowie aus der Perspektive der Gemeinschaften, deren Leben und Lebensunterhalt am unmittelbarsten von dieser Wirtschaft betroffen sind.
P117	Kshirsagar et al. (2021) erläutern auf der Grundlage ihrer Erfahrungen in AI for good (AI4G) Projekten die verschiedenen Aspekte der Zusammenarbeit.	Die Publikation stellt elf Empfehlungen für künftige Projekte zu den Themen Kommunikation, Daten, Modellierung und Auswirkungen vor.

P127	Leavy et al. (2021) schlagen den Ansatz der Datenkuration mit Einbezug der feministischen und Rassismus kritischen Theorien vor.	Die Publikation entwickelt einen ethischen Rahmen für die Datenkuration, um einen Beitrag für KI-Ethik zu leisten und den Schutz von Grund- und Menschenrechten sicherzustellen.
P129	Lee et al. (2019) verknüpfen verschieden Risikomanagement-prozesse und stellen fest, dass es eine Lücke bei den Maßnahmen gibt, die KI-Entwickelnden helfen kann, befangenheitsbezogene Risiken zu erkennen.	Um die Lücke zu schließen, stellt die Publikation eine Methodik und einen Fragebogen zur Identifizierung von Befangenheit vor und veranschaulicht die Methodik anhand eines Anwendungsfalls.
P140	Jo et al. (2019) argumentieren, dass eine neue Spezialisierung innerhalb des ML gebildet werden sollte, die sich auf Methoden zur Datenerhebung und -beschriftung konzentriert.	Die Publikation zeigt auf, wie aus den Archiv- und Bibliothekswissenschaften Ansätze zur Datenerfassung für soziokulturelle Daten aussehen könnte.
P165	Park et al. (2021) konzentrieren sich auf die Repräsentation des Alters und fragen, ob ältere Erwachsene im Verhältnis zur Gesamtbevölkerung in KI-Datensätzen vertreten sind.	Das Ergebnis der Untersuchung ist, dass ältere Erwachsene stark unterrepräsentiert sind.
P170	Posada et al. (2021) untersuchen, wie die Datafizierung, d.h. die Quantifizierung menschlicher und nicht-menschlicher Faktoren in binären Codes, die Identität von Individuen und Gruppen beeinflusst.	Die Publikation verwendet die Analogie der des Buchdrucks in der Frühzeit, um einen Rahmen für das Verständnis des konstitutiven Wandels von Datafizierung zu schaffen.
P184	Schelenz et al. (2021) stellt eine Designlösung für eine diversitätsbewusste Plattform vor.	Sie zeigen den Gestaltungsprozess der diversitätsbewussten Plattform auf und gehen u.a. die Sammlung von Daten zur Entwicklung diversitätssensibler Algorithmen ein.
P192	Sharma et al. (2020) zeigen auf wie ein Idealwelt- Datensatz erstellt werden kann, der alle soziale Gruppen fair darstellt.	Die Datensätze werden um Die synthetische Datenpunkte erstellt.
P204	Leavy et al. (2021) geben einen Überblick über Gerechtigkeit in Daten und untersuchen, ob die unvermeidliche Voreingenommenheit in KI-Trainingsdaten tatsächlich für das Gemeinwohl genutzt werden kann.	Sie kommen zum Ergebnis, dass intersektionale, feministische und Rassismus kritische Theorien bei der Datenerfassung berücksichtigt werden und die Betroffenen in den in KI-Entwicklungen einbezogen werden sollten.
P209	Tommasi et al. (2021) definieren operative Schritte für eine faire Zertifizierung von KI.	Die Publikation schlägt Maßnahmen für aktuelle Lücken vor, die sie im Bereich Daten und Algorithmen identifiziert hat.
P212	Tomasev et al. (2021) untersuchen die Belange von Queer Communities im Bereich der algorithmischen Fairness.	Dieser Beitrag unterstreicht die Bedeutung der Entwicklung neuer Ansätze für algorithmische Fairness, die sich von der vorherrschenden Annahme beobachteter Merkmale lösen soll.

P219	Vredenburgh (2022) untersucht die vernachlässigte normative Dimension der algorithmischen Intransparenz am Arbeitsplatz und auf dem Arbeitsmarkt untersucht.	Die Publikation zeigt als Ergebnis verschiedene Entfremdungsdimensionen in der KI-gesteuerten Arbeitswelt auf.
P222	Whittlestone et al. (2019) zeigen anhand von Vergleichen zwischen KI-Ethik und Bioethik einige der Grenzen von Grundsätzen auf.	Die Publikation erörtert einige spezifische Arten von Spannungen in der KI-Ethik und welche Prozesse erforderlich sein könnten, um sie zu lösen.
P238	Zuiderveen Borgesius (2020) bewertet den derzeitigen Rechtsschutz in Europa gegen diskriminierende algorithmische Entscheidungen.	Die Publikation plädiert für sektorspezifische Regulierungen und skizziert einen Ansatz zur Regulierung algorithmischer Entscheidungsfindung.

Eidesstattliche Erklärung

Mit meiner Unterschrift erkläre ich, dass die vorliegende Arbeit selbständig und nur unter Verwendung der im Literaturverzeichnis aufgeführten Quellen erarbeitet worden ist. Die Stellen meiner Arbeit, die dem Wortlaut oder dem Sinn nach anderen Werken entnommen sind, habe ich in jedem Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht. Die Angaben sind für jede einzelne Quelle als Fußnote mit Verweis auf die Quelle aufgeführt. Dasselbe gilt sinngemäß für Tabellen, Karten und Abbildungen, auch solche, die aus Internetquellen stammen.

Düsseldorf, den 20. Juli 2022 [redacted]

Ort, Datum | Unterschrift