Bias in KI-Systemen: Risiken und Lösungsansätze zur Minimierung von Verzerrungen

Bachelorstudium Informatik

Abgabe: [XX.XX.XXXX]

Inhaltsübersicht

1. Einleitung

Künstliche Intelligenz (KI) spielt eine wachsende Rolle in vielen Bereichen, von der Automatisierung von Büroprozessen über die Analyse medizinischer Bilder bis hin zur Entscheidungsunterstützung von Richtern und Politikern. Allerdings stehen durch die Anwendung dieser technischen Intelligenz zentrale Fragen im Raum. Wie objektiv sind die durch KI getroffenen Entscheidungen tatsächlich? Spiegeln sich die menschlichen Vorurteile nicht auch in den Computerprogrammen wider?

Vor allem der Umgang mit systematischen Verzerrungen, sogenannten Bias, ist ein Thema, da mit den Entscheidungen von KI-Systemen Konsequenzen einhergehen können. Bei einem Bias handelt es sich um eine strukturelle Verzerrung in einer Datenstichprobe oder einem Algorithmus aufgrund des Designs oder der Konstruktion. Häufig liegt eine verzerrte Stichprobe, ein fehlerhafter Algorithmus oder ein gewählter Parameter vor. Dadurch werden Vorurteile und Ungerechtigkeiten manifestiert. Die Herausforderung ist es, Verzerrungen in KI-Systemen nachweisen zu können, um diese einzudämmen. Die ethische, soziale und rechtliche Bewertung solcher Systeme ist unerlässlich. So geht es beim Umgang mit Verzerrungen und Ungleichheiten in KI-Systemen um Fragen der Transparenz, Verantwortlichkeit, Datenschutz, Schutz von gefährdeten Personen oder Gruppen und der Fairness.

In der vorliegenden Arbeit sollen diese systematischen Bias-Effekte in KI-Systemen näher betrachtet werden, und es soll aufgezeigt werden, wie ein solcher Bias detektiert und minimiert werden kann. Im Fokus steht die Fragestellung, woher Bias in KI-Systemen kommt, wie er sich auf die Entscheidungsprozesse auswirkt und was getan werden kann, um Verzerrungen zu minimieren. Dies wird in Form einer umfassenden Literaturanalyse (quantitativ und qualitativ) untersucht.

Der folgende Überblick gibt einen Einblick in den Aufbau der vorliegenden Hausarbeit: Im zweiten Kapitel werden zunächst die theoretischen Grundlagen in Bezug auf Bias erläutert. Das dritte Kapitel beschäftigt sich mit den Ursachen und Quellen von Bias, die in drei verschiedene Gruppen unterteilt werden (datenbasiert, algorithmisch und durch menschlichen Input entstanden). Anschließend, im vierten Kapitel, werden die Risiken und Auswirkungen von Bias beschrieben, indem die Verzerrungen in unterschiedlichen Bereichen der KI analysiert werden. Im fünften Kapitel werden unterschiedliche Lösungsansätze (technische und regulatorische Lösungen) im Umgang mit Bias präsentiert.

Die wichtigsten Erkenntnisse werden im sechsten Kapitel zusammengefasst, und es werden ein Ausblick auf die kommenden Herausforderungen und Handlungsempfehlungen formuliert.

2. Grundlagen von Bias in KI-Systemen

Bias in KI-Systemen bezeichnet systematische Verzerrungen, welche zu unfairen oder diskriminierenden Ergebnissen führen, da sie auf historischen Diskriminierungsmustern basieren und diese fortführen (Rönnecke, 2024, S. 1). Algorithmen entscheiden und sind derweil unfair und diskriminierend gegenüber gesellschaftlichen Gruppen. Auch werden durch KI-Systeme historische Diskriminierungen automatisch und unkontrolliert reproduziert (Rönnecke, 2024, S. 1). Bias in KI-Systemen findet z. B. bei der polizeilichen Prognose von Straftaten und der automatisierten Jobsuche statt und stellt ein Problem der sozialen Gerechtigkeit dar (Rönnecke, 2024, S. 1).

Wann systematische Verzerrungen nicht zu einem Problem führen, wenn sie auf dem gesellschaftlichen Status quo beruhen, verfestigt sich der Status quo durch systematische Verzerrung. Die Analyse und Überprüfung der Daten, auf denen KI-Systeme basieren, erfordert neue Methoden, da bestehende Vorurteile oft nicht erkannt werden können (Rönnecke, 2024, S. 2).

Die wichtigsten Ursachen für Bias in KI-Systemen bestehen zum einen in Fehlern bei der Datenbeschaffung und zum anderen in algorithmischen und menschlichen Einflüssen. Technische Verzerrungen werden insbesondere durch unzureichende Datenrepräsentation verursacht. So werden zum Beispiel Menschen mit Behinderung, alte und zu benachteiligenden Gruppen zählende Menschen unterrepräsentiert (Rönnecke, 2024, S. 1). Menschen mit dunkler Hautfarbe werden vermehrt falsch erkannt als Menschen mit heller Hautfarbe (Ferrara, 2023, S. 3). Der Grund hierfür liegt zum einen in der Datenrepräsentation und zum anderen in technischen Mängeln. Bei gesellschaftlichen Verzerrungen kommt es in den verwendeten Daten zur Korrelation von geschützten Merkmalen, wie zum Beispiel Geschlecht, Hautfarbe und ethnische Herkunft mit anderen Merkmalen, sodass bestehende Vorurteile durch Algorithmen reproduziert werden (Rönnecke, 2024, S. 1). Menschliche Einflüsse finden z. B. in Form unbewusster Vorurteile der Entwickler*innen statt, die sich auf das Design von Algorithmen auswirken (Ferrara, 2023, S. 3). Die beiden Verzerrungsarten bedingen und verstärken sich gegenseitig und

erschweren die Minimierung von Bias in KI-Systemen.

Ein konkretes Beispiel sind KI-gestützte Gesichtserkennungstechnologien, welche bei Menschen mit dunkler Hautfarbe deutlich höhere Fehlerraten aufweisen (Ferrara, 2023, S. 3). Ein weiteres Beispiel ist das COMPAS-System, ein US-amerikanisches System für die Prognose von Rückfallrisiken, bei dem Studien belegen, dass afroamerikanische Angeklagte unabhängig von individuellen Vorstrafen systematisch zu Unrecht als risikobehafteter prognostiziert werden als weiße Angeklagte (Ferrara, 2023, S. 3; Yapo & Weiss, 2018, S. 3). Die Ursachen für derartige Verzerrungen können bei der Datenauswahl und Entwicklungsmethodik von KI-Systemen verortet werden und erfordern eine umfassende Analyse der KI-Entwicklungspraktiken und verbesserte Ethikstandards in der Forschung. Das bedeutet, dass die in den Beispielen automatisiert durchgeführten Diskriminierungen nicht auf einzelne Personen, sondern auf ganze gesellschaftliche Gruppen und ganze Regionen zutreffen.

Darüber hinaus kann eine algorithmische Reproduktion der Norm dazu führen, dass bestimmte stereotype Rollen in der Gesellschaft nicht nur erhalten werden, sondern zusätzlich verstärkt werden (Ferrara, 2023, S. 4). So liefern zum Beispiel die KI-Bildgenerierungstechnologien DALL-E und Midjourney beim Prompt "CEO" ausschließlich Männer (Ferrara, 2023, S. 4). Bei Job-Suchportalen zum Beispiel sehen männliche Interessierte die Suchergebnisse für qualifizierte Jobangebote deutlich häufiger angezeigt als weiblich gelesene Profile (Yapo & Weiss, 2018, S. 2). KI-Systeme können also zu einer Verfestigung von bestehenden Stereotypen über Geschlecht und Geschlechterrollen führen (Smith & Rustagi, 2020, S. 42). Diese Befunde legen nahe, dass KI-Systeme, wenn nicht entsprechende Methoden zur Korrektur von Bias zum Einsatz kommen, bestehende soziale Diskriminierungen und Ungleichheiten weiter zementieren können.

Die Schwierigkeiten bei der Identifikation und Korrektur von Bias in KI-Systemen bestehen besonders in dem Umstand, dass oftmals Transparenz bezüglich der zugrunde liegenden Algorithmen von KI-Systemen nicht vorliegt (Rönnecke, 2024, S. 1). Wissenschaftler*innen fordern Maßnahmen wie die Offenlegung der Trainings- und Testdaten sowie Audits, um Schwachstellen und Diskriminierungen aufzudecken (Rönnecke, 2024, S. 2; Smith & Rustagi, 2020, S. 9). Ob dies umsetzbar ist, darüber entscheidet der weitere Forschungs- und Entwicklungsfortschritt.

3. Quellen und Ursachen von Bias

Die Ursachen von Bias in KI sind vielfältig. So reichen sie von datenbasierten über algorithmische bis hin zu menschlichen Einflüssen. Unzureichende Repräsentation, historische Diskriminierung sowie individuelle Vorurteile sind nur einige der Einflüsse auf die Datenqualität. In diesem Abschnitt werden diese zusammenfassend betrachtet, um nachvollziehen zu können, wo die unterschiedlichen Ursachen von Bias zu finden sind und um Gegenmaßnahmen zur Minimierung der negativen Auswirkungen dieser für die anschließende Bearbeitung der Fragestellung in der Arbeit ableiten zu können.

3.1 Datenbasierte Verzerrungen

Datenbasierte Verzerrungen bilden eine wesentliche Ursache für Bias in KI-Systemen und führen zur systematischen Benachteiligung von Bevölkerungsgruppen. Ein Hauptgrund für Diskriminierungen ist die mangelnde Repräsentation bestimmter Gruppen in den Trainingsdaten von KI-Systemen. Hierdurch werden systematische Ungleichheiten im Ergebnis verankert. Durch Unterrepräsentationen von minorisierten Gruppen, wie beispielsweise Menschen mit einer Behinderung oder Menschen mit nicht-europäischer Herkunft, erkennt eine Gesichtserkennung diese nicht oder erkennt sie nur schlechter (Rönnecke, 2024, S. 1; Rathgeb, 2023, S. 20). Die Frage, warum solche Unterrepräsentationen in Datensätzen erst relativ spät erkannt und korrigiert werden, wird durch die mangelhafte Datenexploration, das heißt fehlende Kontrollschritte und Überprüfungen bei der Datenvorbereitung, sowie einen systematischen Erfassungsfehler, also eine selektive Versäumnis in der Erfassung, begründet.

Ein weiteres Problem ergibt sich durch die nahezu kritische Übernahme von Diskriminierungen in den Datensätzen. KI-Systeme haben keine Möglichkeit, die historische Gültigkeit dieser Diskriminierungen selbst zu analysieren. Sie reproduzieren gesellschaftliche Ungleichheiten (Rönnecke, 2024, S. 1). So konnten Datensätze aus älteren Anwendungen zur Kreditvergabe und Beschäftigung, die negative Muster über Gruppen in der Gesellschaft widerspiegelten, weiterverwendet werden. Damit werden auch diskriminierende Vorurteile weitergereicht. Eine Herausforderung ist daher, historische Verzerrungen aus den Datensätzen zu identifizieren und diese auszuräumen. Eine Korrektur dieser Verzerrungen in dem Vorfeld der Modellentwicklung ist allerdings nicht trivial, da es hierfür keine etablierten Korrekturverfahren gibt, die alle Risiken abdecken. Das Ausmaß der

Korrektur ist daher schwierig zu bestimmen. Zudem müsste diese repräsentativ sein, was bei großen Datenmengen enormen Aufwand bedeutet.

Das ungenügende Vorhandensein von Gruppen in den Daten führt zudem dazu, dass die Genauigkeit der Ergebnisse, basierend auf Optimierungen in der dominanten Gruppe, überproportional beeinflusst wird. Denn die Norm basiert auf der quantitativ vorherrschenden Gruppe. Minderheiten können damit nicht exakt berücksichtigt werden. Hierdurch können sich beispielsweise Entscheidungsgrundlagen in bestimmte Richtungen verschieben (Rathgeb, 2023, S. 20). Eine Herausforderung ist es daher, die Komplexität von sozialen Gruppen abzubilden. Durch geeignete Konzepte kann eine Auswahlstrategie entwickelt werden, um möglichst repräsentative Datensätze zu generieren. Es fehlt jedoch noch oftmals an der Datenerfassung. Die Data Diversity Selection illustriert einen blinden Fleck in der Datenstrategie einiger Unternehmen.

Die mangelhafte Ausgangsdatenqualität potenziert das Problem der Unterrepräsentation. Qualitativ schlechte Daten, wie Bilder mit schlechter Auflösung, beschränken beispielsweise Personen mit dunklerer Hautfarbe oder Kinder. Daher werden diese vom Algorithmus auch häufiger nicht erkannt oder falsch klassifiziert (Rathgeb, 2023, S. 19; Anslinger, 2021, S. 4). In der Datenaufbereitung bilden sich somit technische und strukturelle Mängel ab, die die zuvor genannten Verzerrungen in den Daten weiter ausweiten. Dadurch, dass bestimmte Ausprägungen im Trainingsdatensatz unterrepräsentiert sind, entstehen negative Muster im Algorithmus. Es erfordert daher ein nachträgliches Anpassen des Modells mittels Data Augmentation. Dies macht eine starke Sensibilisierung erforderlich.

Ein anderes Problem ist die Abbildung gesellschaftlicher Diskriminierung in statistischen Korrelationen in den Daten. Dadurch können sich verzerrte Zusammenhänge der Daten bilden, die nicht kausal bedingt sind. Geschützte Merkmale, wie Alter, Geschlecht oder Herkunft, korrelieren untereinander und lassen somit automatische Benachteiligungen im Entscheidungsergebnis zu (Aysolmaz et al., 2020, S. 2; Rathgeb, 2023, S. 18; Anslinger, 2021, S. 5). So ergab eine Studie zu Gesichtserkennungstechnologien, dass sich bei afrikanisch- und asiatisch-stämmigen Personen die Fehlerrate gegenüber europäisch-stämmigen Personen in bestimmten Gesichtserkennungssystemen erhöht. Zusätzlich hatten Frauen ein erhöhtes Risiko, durch den Falschakzeptanzfehler fälschlicherweise erkannt zu werden (Rathgeb, 2023, S. 18). Es stellt sich daher die Frage nach der Ursache. Zum einen sind das technische Mängel in den Systemen oder schlechte Qualitätsdaten. Zum anderen kann ein Algorithmus jedoch immer nur die Güte der Daten reflektieren. Dadurch, dass statistische Korrelationen nicht geprüft wurden, entstehen hier

verzerrte Entscheidungen über Personen.

Ein sehr deutliches Beispiel für datenbasierte Verzerrungen zeigte das amerikanische COMPAS-System. Dieses war ein Programm zur Rückfallprognose von Straftätern. Minderheitengruppen wurden von dem System häufiger als hohes Risiko erkannt als Weiße (Aysolmaz et al., 2020, S. 2). Somit werden verzerrte Risikoprognosen aufgrund von fehlerhaften Daten statistisch sichtbar. Um systematisch Verzerrungen auszugleichen, muss es im System selbst eingebaute Kontrollinstanzen geben. Da Verzerrungen oftmals komplex und schwer zu identifizieren sind, werden diese aber erst mit konkreten Ergebnisbeispielen erkennbar, da erst durch die Verwendung der Ergebnisse Bias deutlich wird. Daher müssen die Modelle auf Bias und Diskriminierung systematisch überprüft werden. Eine kritische Datenbasis ist unerlässlich, da es andernfalls fast unmöglich ist, die Ursachen dieser verzerrten Ergebnisse zu korrigieren.

Die anfälligste Phase für die Entwicklung von datenbasierten Verzerrungen liegt in der Auswahl, Erfassung und Aufbereitung der Daten, da diese Prozesse am komplexesten und anfälligsten für Verzerrungen sind (Aysolmaz et al., 2020, S. 6; Rönnecke, 2024, S. 1). Vorannahmen, Gewohnheiten von Institutionen und gesellschaftliche Wertungen werden genutzt, um Verzerrungen zu generieren. Die fehlende Dokumentation der erhobenen Daten über deren Erhebung, Herkunft und Verwendung (Data Governance) ermöglicht es, nicht ausreichend Transparenz über das in KI-Systemen verwendete Wissen zu erhalten. So kann nicht überprüft werden, welche Daten in welchem Umfang berücksichtigt worden sind (Aysolmaz et al., 2020, S. 6). Ebenfalls ist nicht klar, welche Entscheidungsparameter auf Grundlage der verfügbaren Daten auf welchen Gruppen beruhen und welchen Einfluss dies auf die Prognose über die Gruppe haben könnte. Standardprozeduren und die Einführung eines Standardbegriffes, der die Datengrundlage beschreibt, werden benötigt, um eine Nachvollziehbarkeit und somit Kontrollierbarkeit herzustellen. Ein weiterer innovativer Ansatz ist die Einführung eines interdisziplinären Projektteams, um somit mehr Diversität einzubringen, damit Verzerrungen leichter entdeckt werden können (Aysolmaz et al., 2020, S. 6).

Auch der Umgang mit Transparenz der Daten ist eine Hürde für die Vermeidung von Bias. Es fehlt oft an den verwendeten Daten von Trainingsdaten über Validierungsdaten bis hin zu Testdaten (Rönnecke, 2024, S. 2; Anslinger, 2021, S. 4). Ein geeignetes Konzept, um diese Hürde zu beseitigen, sind sogenannte Model Cards, in denen alle wichtigen Informationen zum Datensatz aufgelistet werden (Rönnecke, 2024, S. 2). Damit können die Anwender eine detaillierte Überprüfung der Modelle machen und einen kritischen Datensatz beisteuern.

3.2 Algorithmische und menschliche Faktoren

Die algorithmischen und menschlichen Faktoren bedingen einander und tragen somit zur Entstehung von Bias in KI-Systemen bei. Ein Teil dieser Problematik ist, dass unbewusst menschliche Vorurteile und Einstellungen in KI-Systeme einfließen. Dies betrifft die Auswahl der Daten, die Modellierung und die Implementierung von Modellen. Dies geschieht, da Entwickler*innen oder Anwender*innen KI-Systeme immer mit eigenen Einstellungen und Werturteilen oder eigenen gesellschaftlichen Stereotypen verbinden (Müller et al., 2020, S. 7). So können während des Lernens soziale oder kulturelle Annahmen eingebracht und somit historische Ungleichheiten fortgeführt werden. Es ist wichtig, diesen Einfluss von persönlichen Einstellungen und Haltungen zu überprüfen, da er oft der Ursprung algorithmischer Bias ist.

Algorithmen verstärken diese menschlichen Verzerrungen, da durch statistische Generalisierungen auf gesellschaftlich benachteiligte Gruppen geschlossen wird. Durch diese Verfahren sind Algorithmen oft die Ursache dafür, dass es zu Diskriminierungen von Bevölkerungsgruppen kommt. Besonders in hoch automatisierten Bereichen werden Urteilsvermögen und Entscheidungsfreiheit der Menschen außer Kraft gesetzt und Entscheidungen werden anhand von Maschinendaten getroffen (Mohabbat Kar et al., 2018, S. 10). KI-Systeme versprechen, Entscheidungen besser auf die individuellen Nutzer*innen abzustimmen als Menschen und sie deutlich effizienter zu gestalten. Es ist allerdings nicht leicht, die Vorannahmen und Stereotype solcher Entscheidungen zu erkennen und zu entfernen. Durch diese Vorgänge werden somit soziale Ungleichheiten reproduziert.

KI-Systeme sind in der Lage, in einigen Fällen individuelle menschliche Fehler zu vermeiden. Sie können jedoch auch zu einer größeren Verbreitung systematischer Vorannahmen und Stereotype führen, wobei viele dieser Voreinstellungen nicht als kritisch erkannt oder korrigiert werden können (Danks & London, 2017, S. 2). Eine empirische Studie zeigt, dass Menschen mit türkisch klingenden Namen beispielsweise deutlich mehr Bewerbungen schreiben müssen als andere Bewerber*innen, bis sie zu Vorstellungsgesprächen eingeladen werden. Das verdeutlicht, wie Vorannahmen und Diskriminierung im algorithmischen System auftreten können (Krüger & Lischka, 2018, S. 17).

Um die Problematik algorithmischer Bias einzudämmen, besteht die Notwendigkeit, technische Standards zu setzen. Es ist wichtig, diese zu etablieren. Eine Möglichkeit, die damit einhergehenden Verzerrungen zu vermindern, ist es, ein Modell transparent zu gestalten. Komplexere Modelle werden oft als Blackbox-Systeme bezeichnet, da der Einfluss der Parameter in Form von Wechselwirkungen untereinander und die Entscheidungsfindung nicht nachvollziehbar ist. Künstliche neuronale Netze funktionieren so, dass selbst Expert*innen die genauen Wechselwirkungen der verschiedenen Parameter nicht genau analysieren können (Müller et al., 2020, S. 7). Können Fehlerquellen und Faktoren, die zu Verzerrungen führen, weder intern noch durch externe Prüfer*innen entdeckt werden, lässt sich schwer überprüfen, wie das System diskriminierende Auswirkungen kontrolliert. Damit einher geht, dass betroffene Nutzer*innen kaum Rechenschaftsplicht durchsetzen können, da sie häufig keinen Einblick in das Vorgehen und die Funktionsweise von algorithmischen Entscheidungen haben. Aus diesen Gründen ist es nötig, technisch kontrollierbare Offenlegungen in allen Entscheidungen und Handlungen einzubinden. Es bedarf regelmäßiger Algorithmen-Audits, um solche Schwachstellen aufzudecken (Mohabbat Kar et al., 2018, S. 21).

Das Verbesserungspotenzial besteht in der Entwicklung erklärbarer KI-Systeme, die auch die Funktionsweise komplexer Modelle nachvollziehbar machen und dokumentieren. Eine wichtige Komponente ist das Bewusstsein der Programmierer*innen für die Auswahl und Gewichtung der Parameter (Müller et al., 2020, S. 7). Diese wirken sich auf die Systementscheidungen aus, sodass selbst kleine Änderungen große und systematische Auswirkungen haben können. Werden beispielsweise nur Trainingsdaten aus einer Stadt wie Pittsburgh verwendet, wird ein KI-System die Normen der Stadt als allgemeingültig ansehen. In anderen Kontexten oder Gebieten können die Ergebnisse deutlich schlechter oder sogar falsch sein. Das System kann somit auf schlechte oder unzureichende Informationen überangepasst sein und es kann zu Fehlentscheidungen in Bereichen wie Gesundheit oder Recht kommen (Danks & London, 2017, S. 3). Der Einbau von zum Beispiel Glättungs- oder Regularisierungsparametern kann dazu beitragen, dass ein System weniger stark auf falsche Daten angepasst wird und Anomalien nicht mit in Betracht gezogen werden (Danks & London, 2017, S. 3).

Eine weitere Hauptursache für algorithmische Verzerrung ist eine schlechte Vielfalt an Trainingsdaten. Um sicherzustellen, dass gesellschaftlich benachteiligende Muster in das System einfließen und im Modell verankert werden, muss die Validierung der Parameter stetig weitergeführt werden. Die Vielfalt der Trainingsdaten kann verbessert werden, indem das Datensample nicht nur von Nutzer*innen stammt, die vorrangig im Internet aktiv sind,

sondern von Menschen, die aus unterschiedlichsten Gründen wenig bis gar nicht aktiv im World Wide Web sind. Eine schlechte Datenqualität erhöht somit die Gefahr algorithmischer Verzerrung. Diese kann zu falschen Entscheidungen führen, da die Wahrscheinlichkeit, dass sie durch unvollständige und fehlerhafte Informationen beeinflusst werden, steigt. Diese unvollständigen oder fehlerhaften Informationen werden dazu verwendet, Schlussfolgerungen auf andere zu ziehen, beispielsweise aufgrund der Herkunft oder des Alters von Nutzer*innen. So können sie, wenn sie auf einen falschen Datensatz trainiert sind, systematische Diskriminierungen hervorrufen. Das System ist jedoch in der Lage, einzelne menschliche Fehler zu reduzieren, ohne dass aber gesamtgesellschaftliche Vorannahmen abgebaut oder vermieden werden können (Mohabbat Kar et al., 2018, S. 10). In New York verhinderte der ADM nachweislich Benachteiligungen bei der Vergabe von Schulplätzen. Der Einsatz des Systems COMPAS in der Strafjustiz führte allerdings zur Reproduktion der Benachteiligung Schwarzer Angeklagter (Krüger & Lischka, 2018, S. 15, S. 18). Da KI-Entscheidungen als objektiver eingestuft werden als Entscheidungen von Menschen, kann es dazu führen, dass Menschen bei algorithmischen Entscheidungen übersehen, dass die Ergebnisse aus voreingenommenen Daten oder aus Vorannahmen hervorgehen.

Aus diesen Gründen sollte der Einsatz technischer Kontrollmechanismen und menschlicher Kontrolle im Zusammenspiel erfolgen. Da ein KI-System in erster Linie auf die Abbildung von Einzelfällen und nicht auf das Aufdecken gesamter gesellschaftlicher Stereotype trainiert wird, kann menschliche Kontrolle dazu beitragen, die Systementscheidungen durch gesamtgesellschaftliche Bewertungen zu erweitern und vor Diskriminierung zu schützen. Es wäre beispielsweise denkbar, die Überprüfung und Implementierung algorithmischer Entscheidungen durch interdisziplinäre Prüfgruppen durchzuführen.

4. Risiken und Auswirkungen von Bias in KI-Systemen

Die Risiken und Auswirkungen von Bias in KI stellen eine zentrale Herausforderung in Hinblick auf die Integration von KI-Systemen in unterschiedliche Gesellschaftsbereiche dar. So kann es bei einer KI-gestützten Entscheidung zu diskriminierenden Ergebnissen kommen, sobald die KI-Trainingsdaten historische Vorurteile widerspiegeln. Die Fehlerrate bei der Gesichtserkennung, vor allem bei dunkleren Hautfarben, führt zu ungerechten

Verdächtigungen. Dadurch kann es bei sicherheitskritischen Anwendungen zu fatalen Folgen kommen. So hat das COMPAS-System der Strafjustiz afroamerikanische Angeklagte doppelt so häufig fälschlicherweise als rückfallgefährdet eingestuft (Ferrara, 2023, S. 3; Krüger & Lischka, 2018, S. 17; Orwat, 2019, S. 27).

Durch die Automatisierung von Entscheidungen durch KI ist die Gefahr gegeben, dass stereotypes und diskriminierendes Denken verstärkt wird. So müssen Menschen mit ausländisch klingenden Namen mehr Bewerbungen als Personen mit deutschen Namen verfassen, um ein Vorstellungsgespräch zu erhalten (Krüger & Lischka, 2018, S. 17).

Es besteht das Risiko der Verwendung von Ersatzinformationen in KI. So werden Merkmale wie Alter, Herkunft oder Geschlecht verwendet, um Entscheidungen zu generieren. Auch wenn die Merkmale nicht explizit in der KI enthalten sind, kann es sein, dass sie implizit über Stellvertreter oder über die statistische Korrelation der Trainingsdaten enthalten sind und sich daraus Vorurteile und Stereotype ableiten.

In KI-Systemen werden bestehende Verzerrungen nicht korrigiert. Solange es keine Kontrollmechanismen gibt, die überprüfen, ob der Datensatz verzerrt ist oder nicht, schreiben KI-Systeme Ungleichheiten einfach fort (Rönnecke, 2024, S. 1-2).

So kann eine niedrige Fehlerrate bei der Gesichtserkennung mit einem hohen gesellschaftlichen Schaden verbunden sein. Wenn es zum Beispiel in großen öffentlichen Räumen zu Fehlalarmen bei der Identifizierung von „gesuchten“ Personen kommt und eine Fehlerrate von einem Prozent in einem bevölkerungsstarken Gebiet herrscht, so werden täglich tausende Menschen fälschlicherweise durchsucht (Krüger & Lischka, 2018, S. 17).

Ein weiteres Risiko besteht in der Verstärkung von Geschlechterstereotypen durch KI. Wenn zum Beispiel Text-zu-Bild-KIs Bilder von Führungskräften erstellen oder Bilder über die Suche „Führungskraft“ abgerufen werden, werden vorrangig Männer abgebildet. So zeigen sie, dass etablierte Vorurteile in ihren Algorithmen repräsentiert werden (Ferrara, 2023, S. 4; Orwat, 2019, S. 36).

Das bedeutet, dass bei vielen KI-gestützten Anwendungen die Reproduktion und Verstärkung von Stereotypen vorangetrieben wird, auch wenn keine Geschlechterzuweisung durch den Nutzenden vorgenommen wird. So wird es dem Nutzer im Internet beispielsweise öfter passieren, dass auf seiner Website Reklame für eine gut bezahlte Jobstelle für Männer gezeigt wird (Ferrara, 2023, S. 4).

Ein weiteres Beispiel für eine algorithmische Reproduktion von Vorurteilen ist die automatische Personalauswahl durch KI, bei der Menschen aus gesellschaftlichen Minderheiten diskriminiert werden (Orwat, 2019, S. 36). So führen verzerrte oder unvollständige Datensätze zu einem blinden Vertrauen und zu einem falsch kalibrierten System (Ferrara, 2023, S. 4).

Verstärkte Stereotypen und Verzerrungen führen zu einer verringerten Chancengleichheit, dadurch auch zu einer niedrigeren Integration in der Gesellschaft und so auch zu einem höheren Risiko, bei Ressourcen dauerhaft ausgegrenzt zu werden (Orwat, 2019, S. 36).

Konkrete Beispiele von Fehlern in der Anwendung von KI in Bereichen wie Bildung, Medizin, etc. zeigen die realen Risiken für die gesellschaftliche Integration (Ferrara, 2023, S. 3, 12; Krüger & Lischka, 2018, S. 15, 18).

So wurde in New York KI genutzt, um die Schüler*innen an Schulen zuzuordnen. Dadurch konnten die Verzerrungen und Nachteile, denen diese Schüler*innen gegenübergestellt wurden, um fast 50 % reduziert werden (Krüger & Lischka, 2018, S. 15).

Es wurde ein KI-System für die Diagnose von Krankheiten in Krankenhäusern eingesetzt, um Sterblichkeit zu prognostizieren. Afroamerikaner*innen hatten im Gegensatz zu hellhäutigen Amerikaner*innen eine niedrigere Sterblichkeitsprognose als die tatsächlich auftretende Sterblichkeit (Ferrara, 2023, S. 3).

Wie eingangs in Kapitel 1.3.1 beschrieben, gibt es Risiken durch Verzerrungen, wenn Entscheidungen durch Algorithmen beeinflusst werden. So kann es durch die Verzerrungen zu einer falschen Prognose des Rückfallrisikos kommen. Schwarze Angeklagte werden durch das System COMPAS für ein doppelt so hohes Rückfallrisiko als weiße Angeklagte geschätzt (Krüger & Lischka, 2018, S. 18).

In der automatischen Rekrutierung über Jobseiten oder in Bewerbungen für Lehrstellen in Österreich stellen Menschen mit ausländischem Namen einen Nachteil dar, bei der Auswahl in eine höhere Schule zu kommen. Dies zeigt, dass KI auch einen strukturellen Nachteil für einen gleichmäßigen und gleichberechtigten Zugang zur Bildung generiert (Krüger & Lischka, 2018, S. 15).

Durch die automatisierte Reproduktion der sozialen Ordnung wird eine technokratische

Fortschreibung der Machtstrukturen unterstützt, wenn KI-Systeme nicht kontrolliert werden (Orwat, 2019, S. 27).

Es gibt ein großes Risiko, dass die Daten und die Entscheidungsfindung schwer verständlich und damit intransparent und nicht kontrollierbar sind.

Das Risiko der Intransparenz birgt große Risiken für die KI-Ethik, weil die Nachvollziehbarkeit und Verantwortung durch sogenannte „Blackbox“-Modelle untergraben wird (Heesen et al., 2021, S. 4).

Je weniger transparent ein System ist, desto schwieriger ist es für andere (Nutzer*innen, Unternehmen, staatliche Behörden) zu verstehen, was es tut, was seine Wirkungen sind und ob es sich an geltende Gesetze hält (Orwat, 2019, S. 22).

Wird nicht offen gezeigt, wie das KI-System entwickelt wurde oder wie die eingesetzten Algorithmen funktionieren, ist es sehr schwierig, etwaige Fehler, Risiken oder Vorurteile des Systems zu entdecken (Krüger & Lischka, 2018, S. 10).

Diese Art des Risikos ist im strafrechtlichen Kontext von Bedeutung, aber auch in anderen Kontexten (z. B. Gesundheitswesen, etc.).

Durch die fehlende Offenlegung ist auch ein höheres Risiko vorhanden, KI-Systeme nicht oder nur bedingt anzuerkennen und ihnen zu misstrauen (Orwat, 2019, S. 22; Heesen et al., 2021, S. 4).

Es gibt auch das Risiko, dass Personen und Institutionen KI-Systemen unkritisch blind vertrauen. Die Mehrheit der Bevölkerung ist nicht in der Lage, die Auswirkungen von KI zu bewerten. Es gibt Untersuchungen, die besagen, dass es nur fünf Prozent der deutschen Bevölkerung gelingt, KI-Systeme kritisch zu beurteilen (Melcher, 2025, S. 1).

Insgesamt zeigen sich Risiken und Auswirkungen von Bias in KI-Systemen auf vielen gesellschaftlichen Ebenen.

5. Lösungsansätze zur Minimierung von Bias

Zur Minimierung von Bias in KI-Systemen sind technische sowie regulative Aspekte zu berücksichtigen. Mithilfe von speziellen Methoden und Verfahren, die Fairness fördern, und durch die Vorgabe bestimmter Gesetze zur Minimierung von Diskriminierung kann der Bias von KI-Systemen reduziert werden. Hierbei wird deutlich, dass verschiedene Maßnahmen gemeinsam zu dem Erfolg des Abbaus gesellschaftlicher Ungleichheiten durch Technologie beitragen.

5.1 Technische Methoden

Die technischen Methoden zur Bias-Minimierung beschäftigen sich vor allem mit der Entwicklung und Anwendung spezifischer mathematischer Methoden sowie organisatorischer Verfahren zur Identifizierung und Verminderung systematischer Verzerrungen. In diesem Zusammenhang spielt vor allem die Entwicklung von Fairness-Algorithmen auf Basis mathematischer Metriken wie „Demographic Parity“ und „Equal Opportunity“ eine entscheidende Rolle. Ziel ist es, marginalisierte Gruppen während der Trainingsphase durch algorithmische Änderungen zu schützen (Ferrara, 2023, S. 3). In der Anwendung lassen sich dadurch teils positive Effekte erzielen. Eine Herausforderung liegt jedoch in der situationsabhängigen Auswahl geeigneter Fairness-Metriken. Eine einseitige Optimierung einer Kennzahl kann wiederum negative Auswirkungen in anderen Bereichen, wie eine reduzierte Genauigkeit für bestimmte Gruppen, haben. Aus diesem Grund ist eine kritische Evaluierung des Trade-offs zwischen Fairness und Genauigkeit erforderlich. Einige Forschungsinitiativen sprechen sich daher auch für dynamische, auf Feedback basierende Fairness-Metriken aus, die die Anpassung der Anforderungen an Algorithmen in Abhängigkeit von gesellschaftlichen Trends ermöglichen.

Technische Maßnahmen zur Bias-Verminderung durch Justierungsmechanismen wie „reweighting“ und „adversarial debiasing“ können ebenfalls zur Minimierung von Verzerrungen beitragen. Beim „reweighting“ werden die Gewichtungen der Datensätze im Trainingsprozess angepasst, sodass unterrepräsentierte Gruppen stärker in den Lernprozess einfließen. Im Gegensatz dazu lernt beim „adversarial debiasing“ ein zweites Modell dazu, die Einflüsse diskriminierender Merkmale auf Entscheidungsfindung zu minimieren. Durch empirische Untersuchungen zeigt sich, dass in Anwendungsbereichen, in denen besonders sensibles Material verarbeitet wird, die Wirksamkeit für unterrepräsentierte

Gruppen mit diesen Methoden erhöht werden konnte (Rathgeb, 2023, S. 18; Ferrara, 2023, S. 3). Auch hier werden jedoch weitere Kompromisse bei der Gesamteffizienz eingegangen, die eine stärkere Interdisziplinarität und damit ethische und rechtliche Berücksichtigung in technischen Fragestellungen bedingen.

Um Verzerrungen zu minimieren, können des Weiteren geeignete Methoden zur Datensatzvorverarbeitung eingesetzt werden. Darunter fallen zum Beispiel Data-Cleaning-Techniken, „sampling"-Methoden, Data Augmentation und synthetische Datengenerierung, mit denen eine gezielte Anhebung des Repräsentationsgrad unterrepräsentierter Gruppen beziehungsweise die Erzeugung synthetischer Daten einhergeht (Rathgeb, 2023, S. 18; Aysolmaz et al., 2020, S. 6). Daten können hierbei auch nach dem Zufallsprinzip ergänzt, gelöscht, umgeschrieben oder gefälscht werden, um mögliche Vorurteile im Data-Set zu unterdrücken und ein stabileres Training zu ermöglichen. Dadurch lassen sich beispielsweise in dem Bereich der Gesichtserkennung die Fehlerraten älterer Menschen und Kinder reduzieren (Rathgeb, 2023, S. 18). Hierbei ist jedoch Vorsicht geboten, denn beim Generieren von synthetischen Daten können neue Verzerrungen eingeführt werden. Aus diesem Grund muss sichergestellt werden, dass synthetische Daten statistisch valide sind, indem sie einem strengen Kontrollverfahren unterzogen werden. Eine kontinuierliche Evaluation der Datensätze ist für einen korrekten Einsatz essenziell.

Transparente Dokumentation und Auditierung sind ein weiterer wichtiger Baustein bei der Minimierung von Bias. Mit Verfahren wie „Model Cards" sowie Testprotokollen und regelmäßigen Algorithmen-Audits werden Daten besser nachvollziehbar gemacht und die Überprüfbarkeit von Diskriminierungen erleichtert (Rebstadt, 2023, S. 7). Die Transparenz von Datenquellen und Datenverarbeitung ist essenziell, um Diskriminierungen kontrollieren zu können. Dies belegen Umfragen der US-amerikanischen Firma Element AI, wonach 79 % der Befragten einen erhöhten Wert auf die Umsetzbarkeit von Transparenz- und Fairnessansätzen legen (Rebstadt, 2023, S. 7). Eine mögliche Schwäche dieser Ansätze liegt darin, dass nur dokumentiert wird, wie, aber noch nicht, was verändert werden muss. Um wirksam gegen Diskriminierungen vorgehen zu können, ist daher ein Monitoring mit verpflichtenden Kontroll- und Korrekturmechanismen sowie klarer Verantwortungszuteilung erforderlich. Um dies zu gewährleisten, können zum Beispiel interaktive Audit-Plattformen eingesetzt werden, mit denen im Fall einer möglichen Diskriminierung Kontrollmechanismen und Handlungsempfehlungen aus der ethischen Bewertung verankert werden können. Eine Weiterentwicklung solcher interaktiven Plattformen ermöglicht es, bereits vor dem Systemeinsatz proaktiv auf Bias zu prüfen und zu reagieren.

Außer technischen Kontrollmechanismen kann die Sensibilisierung von Entwickler*innen zu einem höheren Problembewusstsein für Verzerrungen führen. Diese Bewusstseinsschärfung lässt sich zum Beispiel durch Trainingsprogramme, Workshops zum Thema Ethik oder eine bessere interdisziplinäre Kommunikation im Entwicklungsprozess erzielen (Busse et al., 2023, S. 7). In besonders komplexen Anwendungsbereichen, zum Beispiel in der Verarbeitung von Sprache und Text, sollte die Identifizierung potenzieller Bias-Quellen im Entwicklungsprozess durch Kontrollmechanismen wie Peer Reviews oder Bias-Impact-Assessments gesichert werden. Durch die Kombination technischer, sozio-technischer sowie organisatorischer Maßnahmen, einheitliche Ethik-Standards in unterschiedlichen Abteilungen und die Interdisziplinarität der Standards sinkt die Wahrscheinlichkeit von diskriminierenden Systemausgängen, die in einigen Studien identifiziert wurden. Zu den Beispielen für die aktive Integration von Gruppen, die besonders von Diskriminierungen durch KI-Systeme betroffen sind, zählen die Mitsprachemöglichkeiten von Betroffenen bei der Entwicklung sowie dem Testen von Systemen, sodass sie zu Expert*innen in der Identifizierung und Verminderung von Bias gemacht werden.

Um diskriminierende Effekte von KI-Systemen über den Entwicklungszeitraum hinaus zu vermeiden, werden regelmäßiges Monitoring und Feedback in Erwägung gezogen. Dies lässt sich durch die Einrichtung öffentlicher Beschwerde- oder Bewertungsportale ermöglichen, wodurch das Problembewusstsein von Nutzer*innen zu Verbesserungsvorschlägen genutzt werden kann. Ein weiteres zentrales Tool für die Minimierung von Bias ist die Festlegung interdisziplinärer Standards und Richtlinien, in denen technische, ethische und organisatorische Maßnahmen zur Bias-Prävention integriert werden (Aysolmaz et al., 2020, S. 6; Ferrara, 2023, S. 4). Diese Standards sollen eine verbindliche und systematische Anwendung in allen Phasen des Entwicklungsprozesses gewährleisten. Dies erfordert, dass die Anforderungen in regelmäßigen Abständen diskutiert, ergänzt und verbessert werden. Die Analyse bisheriger Forschungsergebnisse legt nahe, dass insbesondere Phasen wie die Datenauswahl und -erhebung die Ausgangsbasis für mögliche Vorurteile bilden. Da Verzerrungen auch in der Datenbereitstellung, durch „missing data“ oder einen unzureichenden Repräsentationsgrad unterrepräsentierter Gruppen auftreten können, ist der transparente Umgang mit Diversitätsaspekten in Dokumentationen ebenso wichtig wie die Diversität bei der Datenbeschaffung. Die Entwicklung solcher Standards ist eine interdisziplinäre Herausforderung. Die enge Zusammenarbeit zwischen verschiedenen Fachbereichen reduziert sogenannte „blinde Flecken“ im Entwicklungsprozess.

Allerdings ist es notwendig, sich kritisch mit der Wirksamkeit solcher Standards

auseinanderzusetzen, da sich viele soziale Normen und Diskriminierungsmuster durch kulturelle oder länderspezifische Ausprägungen auszeichnen, sodass dynamische und für den Anwendungsfall angepasste Standards erforderlich sind. Hierzu werden zum Beispiel sogenannte „Living Documents“ und Ansätze zur Nutzung von Community-Feedback vorgeschlagen, die ein hohes Maß an Anpassungsfähigkeit der Richtlinien ermöglichen sollen. Die beschriebenen technischen Methoden, kombiniert mit organisatorischen Maßnahmen sowie interdisziplinärer Kommunikation, können dazu beitragen, die Komplexität der Bias-Minimierung zu beherrschen und eine Basis für vertrauenswürdige und ethisch akzeptable KI-Systeme zu schaffen.

5.2 Regulatorische Ansätze

Die Regulierung von KI-Systemen dient dem Ziel, Transparenz und Kontrolle in algorithmischen Entscheidungsprozessen herzustellen. Die Einführung staatlicher Prüfstellen trägt dem Bedürfnis vieler Menschen nach mehr Sicherheit Rechnung, da diese den fairen Algorithmen vertrauen (Müller et al., 2020, S. 6). Ebenso dienen diese Einrichtungen dazu, systematische Überprüfungen der KI-Systeme durchzuführen und vor allem die Nichteinhaltung der Antidiskriminierungsvorschriften festzustellen. Dafür überprüfen die Prüfstellen verwendete Trainingsdaten und algorithmische Entscheidungen (Müller et al., 2020, S. 6; Mohabbat Kar et al., 2018, S. 21). Dies birgt jedoch die Herausforderung, adäquate und transparente Prüfkriterien und Auditverfahren zu entwickeln. Die Interdisziplinarität der Prüfgremien und die Veröffentlichung der Auditberichte können den Erfolg dieses Regulierungsansatzes stärken (Mohabbat Kar et al., 2018, S. 21).

Im Zuge der europäischen KI-Verordnung müssen Entwickler*innen alle Trainings-, Validierungs- und Testdatensätze auf Verzerrungen untersuchen (Rönnecke, 2024, S. 2). Hier zeigt sich eine Verbesserung im Vergleich zu vorherigen Verordnungen, da diskriminierende Muster schon vor der Anwendung identifiziert werden können. Von diesem Ansatz sind vor allem Kreditvergaben und Anwendungen im Personalauswahlbereich betroffen. Es wird jedoch infrage gestellt, ob die genutzten Mess- und Analyseverfahren präzise genug sind, die strukturelle Benachteiligung unterrepräsentierter Gruppen zu verhindern (Rönnecke, 2024, S. 2).

Weitere regulatorische Maßnahmen beinhalten zum Beispiel verbindliche Algorithmen-Audits. Solche Algorithmen-Audits beinhalten Prüfungen der Datenqualität,

Fairness von Erfolgskriterien und Vorhersagegenauigkeit (Mohabbat Kar et al., 2018, S. 21). Auch hier sind Datenbereinigungen und Algorithmenkorrekturen Maßnahmen, die im Falle einer Diskriminierung eingesetzt werden können. Die Durchführung und Wirksamkeit solcher Maßnahmen wird von den Prüfressourcen und dem Einblick in die verwendeten Daten beeinflusst. Unternehmen, die aus Geheimhaltungsgründen keinen Dateneinblick gewähren, erschweren eine effiziente Regulierung (Mohabbat Kar et al., 2018, S. 21).

Dass gesetzliche Vorgaben notwendig sind, um Diskriminierung zu verhindern, zeigt das Beispiel des Gesetzes, welches vorschreibt, dass Scoring nicht ausschließlich durch Verwendung der Adressdaten stattfinden darf (Mohabbat Kar et al., 2018, S. 19). Diese Regelung verhindert jedoch nur die explizite Verwendung von Daten zur Kategorisierung von Menschen, wie zum Beispiel hinsichtlich ihrer Wohnortzugehörigkeit. Ob aber die verwendeten Variablen eine verzerrte Gewichtung erfahren und welche Algorithmuskorrekturen in diesem Zusammenhang notwendig sind, liegt oft im Dunklen. Des Weiteren zeigen gesetzliche Regulierungsversuche auch ihr Gefahrenpotenzial, das sie bergen. So können Maßnahmen zwar zur Vorbeugung von Diskriminierung entwickelt werden, diese jedoch durch ihre faktische Anwendung in der Praxis nicht ihre Ziele erfüllen.

In Bezug auf die Verwendung von „Blackbox“-Modellen, die aufgrund ihrer Komplexität und Intransparenz ein hohes Diskriminierungsrisiko haben (Müller et al., 2020, S. 7), ist anzumerken, dass die regulatorischen Anforderungen mittlerweile nicht mehr nur die Offenlegung der Entscheidungslogik, sondern auch die explizite Verwendung erklärbarer KI-Methoden einfordern. Model Cards und die Technische Dokumentation von KI-Systemen tragen zu mehr Transparenz bei, indem sie Betroffenen einen umfassenden Zugang zu den algorithmischen Entscheidungen ermöglichen (Müller et al., 2020, S. 7; Rönnecke, 2024, S. 1). Hier zeigt sich jedoch ein Kritikpunkt: Im Vergleich zu einfachen Regelwerken stellen diese Kontrollmaßnahmen erhebliche zusätzliche Belastungen für KI-Entwickler dar. Dies kann sich vor allem in der Praxis mit der Verfolgung unternehmerischer Ziele und der Produktionseffizienz negativ bemerkbar machen.

Ein vielversprechender Ansatz im Bereich der Regulierung von KI-Systemen findet sich im Zusammenspiel zwischen Transparenz und Rechenschaftspflicht. Das Ausweiten der Rechenschaftspflicht gegenüber den von den Algorithmen betroffenen Menschen ist ein Weg, eine erhöhte Kontrolle über die Algorithmen zu gewährleisten und ihnen gleichzeitig die Möglichkeit zu geben, sich gegen diskriminierende Ergebnisse zur Wehr zu setzen (Rönnecke, 2024, S. 1). Sofern eine Haftung für die algorithmische Entscheidung besteht, wird erwartet, dass Betroffene mit entsprechenden Anfragen an die Verursacher herantreten

und eine transparente Nachvollziehbarkeit der KI-Systeme sowie deren Entscheidungen verlangen. Auch dieser Regulierungsansatz weist kritische Aspekte auf. Da sich die Erwartungen in die Fähigkeit der Unternehmen, Selbstregulierungsinitiativen vorzulegen, bisher nicht erfüllt haben, scheinen gesetzliche Standards für Transparenz zwingend notwendig zu sein (Müller et al., 2020, S. 7; Rönnecke, 2024, S. 1).

Regulatorische Bemühungen auf nationaler Ebene können durch die stetige Globalisierung von Technologien erschwert werden (Antunes et al., 2024, S. 44). So zeigt die Entwicklung einer universellen, interdisziplinären und internationalen Standardpraxis ihre Notwendigkeit, denn es entsteht zum Beispiel eine Lücke im Diskriminierungsschutz, wenn globale Technologieunternehmen von den verschiedenen rechtlichen Standards Gebrauch machen und in Märkte expandieren, die diese nicht oder nur unzureichend regulieren. Studien haben gezeigt, dass Testmetriken, die mit der EU-Gesetzgebung im Hintergrund entstanden sind, auf andere Länder übertragbar sind. Ihre Auswirkungen können jedoch in Bereichen wie der Medizin und des Arbeitsmarktes differieren (Antunes et al., 2024, S. 47). Es muss angestrebt werden, vergleichbare Tools und Best-Practice-Beispiele im gesamten internationalen Kontext einzusetzen (Antunes et al., 2024, S. 47). Dieses Vorgehen ist jedoch nicht ohne Tücken. Durch kulturelle und rechtliche Unterschiede sind flexible Leitbilder notwendig, die die Datenbereinigungspraktiken an die unterschiedlichen Märkte und Anwendungen anpassen.

Internationale Standards können zusätzlich dazu genutzt werden, um zu verhindern, dass ein Algorithmus nur auf den Märkten von Ländern eingesetzt wird, in denen keine Regulierungen existieren. Hier kommt das Prinzip der „Mindestregulierungsstandards“ ins Spiel, welches das Risiko einer globalen Diskriminierung minimiert. Gesetzgebungen sollen einheitliche Standards vorschreiben und die Betreiber und Hersteller dazu verpflichten, diese für alle Anwendungen von KI-Systemen einzuhalten (Antunes et al., 2024, S. 44).

Zusammenfassend ist zu sagen, dass regulatorische Maßnahmen zu einem entscheidenden Werkzeug auf dem Weg zu Transparenz, Nachvollziehbarkeit und Fairness in Bezug auf den algorithmischen Entscheidungsprozess geworden sind.

6. Fazit

Die vorliegende Arbeit hatte es sich zum Ziel gesetzt, das Phänomen des Bias in

KI-Systemen umfassend zu analysieren und Möglichkeiten zur Verringerung diskriminierender Auswirkungen aufzuzeigen. Vor dem Hintergrund der einleitenden Fragestellung, wie es zu systematischen Verzerrungen in datengetriebenen Technologien kommt, wie diese auf gesellschaftliche Bereiche wirken und mit welchen technischen und rechtlichen Ansätzen ihnen begegnet werden kann, wurde dieses Bestreben durch die Auseinandersetzung mit aktueller Forschung erfüllt. Der Zielsetzung, die Ursachen, Risiken und Lösungen von Bias in KI auf interdisziplinärer Ebene darzulegen und zu reflektieren, wurde durch die Zusammenführung zentraler Forschungsstränge und die Zusammenfassung der Implikationen für eine gerechte KI-Entwicklung Nachdruck verliehen.

Im Kern der Arbeit wurde deutlich, dass Bias in KI-Systemen ein komplexes, historisch, technisch und gesellschaftlich bedingtes Phänomen ist. So zeigte sich, dass die Ursachen in datengetriebenen Verzerrungen, wie der Unterrepräsentation von Minderheiten, sowie technischen und menschlichen Faktoren zu finden sind. Die Reproduktion und Verstärkung von Stereotypen und Vorurteilen durch algorithmische Entscheidungen stellt für soziale Gerechtigkeit eine akute Gefahr dar, da durch eine Verlängerung von Benachteiligungen Risiken für das Justizwesen, den Arbeitsmarkt oder das Bildungssystem entstehen. Die Auswertung der Auswirkungen machte darüber hinaus klar, dass die Gefahr dieser Risiken nicht nur in individuellen Fehlentscheidungen, sondern auch in der strukturellen Verfestigung von Verzerrungen besteht. Die fehlende Nachvollziehbarkeit und die niedrige Kontrollierbarkeit komplexer KI-Modelle, welche in der Arbeit herausgestellt wurden, stellen weitere Risiken für einen gerechten Einsatz dieser Technologien dar. Wie die eingehende Beschreibung der technischen und regulatorischen Lösungsansätze gezeigt hat, ist zur Eindämmung des Bias in KI-Systemen eine Kombination von fairen Algorithmen, Dokumentation, Audits und gesetzlichen Vorgaben notwendig, um diskriminierende Auswirkungen möglichst niedrig zu halten. Auch hier zeigt sich allerdings, dass die technische und rechtliche Behebung des Bias in KI-Systemen an Grenzen stößt, sobald man die komplexen Einflüsse von Gesellschaft, Kultur, praktischen Hindernissen und internationalen standardspezifischen Problemen betrachtet.

Aus Sicht der Wissenschaft wird deutlich, dass die Problematiken des Bias und der Fairness von KI-Anwendungen im Fokus der internationalen Forschung liegen und laufend neue Erkenntnisse gewinnen. Die Forschungsergebnisse in dieser Arbeit lassen sich in die zentralen Befunde bezüglich der Komplexität des Bias in KI-Systemen aus führenden wissenschaftlichen Studien einordnen. Gleichzeitig erweitern die Ergebnisse dieser Arbeit die Forschung zum Verständnis über die Ursachen von Verzerrungen und Lösungsansätze, indem im Besonderen auf die Zusammenhänge von technischen, datengetriebenen und

gesellschaftlichen Faktoren eingegangen wird. Die kritische Betrachtung der Grenzen technischer Lösungsansätze und die Verortung der regulatorischen Vorgänge auf internationaler Ebene machen die Arbeit für die aktuelle Forschung in Bezug auf eine interdisziplinäre, internationale Betrachtung wertvoll. Als Beitrag zur Versachlichung der Diskussion um Fairness in KI leistet die vorliegende Arbeit darüber hinaus auch einen wichtigen Schritt für ein besseres Verständnis des Bias-Phänomens.

Dennoch zeigen sich bei der Betrachtung der erläuterten Lösungsansätze Lücken. Die eigenen Forschungsgrenzen ergeben sich vor allem durch die theoretisch-analytische Ausarbeitung, die sich in erster Linie auf ausgesuchte Anwendungsfelder konzentriert hat, wodurch eine tiefer gehende empirische Untersuchung und eine sektorielle Spezifität der Befunde und Herausforderungen nur begrenzt abgedeckt wurden. Forschungsleitende Fragen in der aktuellen Debatte betreffen vor allem die Implementierung von technischen und organisatorischen Fairness-Standards, die praktische Durchführung von Auditverfahren und die gesellschaftliche Mitgestaltung bei der Entwicklung von KI-Anwendungen. In diesen Bereichen zeigt sich zudem der Bedarf an einer stärkeren internationalen Koordination, welche durch die Harmonisierung regulatorischer Vorgänge weiter forciert werden sollte. Die Forschung und die Anwendung von KI-Systemen sollten vor allem im Bereich der interdisziplinären Kooperation, Weiterbildungsmaßnahmen und Gesetze an den technologischen Fortschritt angepasst und damit für eine faire und gerechte Entwicklung von KI-Systemen genutzt werden.

KI-Systeme übernehmen immer mehr Entscheidungen in wichtigen Lebensbereichen und treffen damit oft nachhaltige Entscheidungen über Einzelne oder Gruppen. Die Reflexion über das eigene Lernverhalten während des Arbeitsprozesses und die gewonnenen Erkenntnisse aus dieser Auseinandersetzung haben deutlich gemacht, wie wichtig eine kritische und reflektierte Gestaltung von technischen Systemen ist. Die gewonnenen Einsichten machen mich für die ethischen und rechtlichen Implikationen beim Einsatz von KI-Systemen sensibilisiert. Sie sollen die Reflexion über die Anforderungen von ethischen und rechtlichen Fragestellungen an KI-Anwendungen auch nach dem Arbeitsprozess verstärken. Ich nehme aus dem Arbeitsprozess und den Befunden das Bewusstsein mit, dass interdisziplinärer Dialog über alle Stakeholder-Gruppen hinweg notwendig ist, um zu einer Versachlichung der Debatte über die Gerechtigkeit in KI-Systemen beizutragen. Insgesamt gibt die Arbeit einen Überblick zum aktuellen Forschungsstand, Risiken und Lösungen in Bezug auf Bias in KI-Systemen und zeigt damit die nächsten Schritte für zukünftige Forschungsarbeiten zum Thema auf.

Literaturverzeichnis

Anslinger, J. (2021). Faire KI - (wie) geht das? IFZ - Interdisziplinäres Forschungszentrum für Technik, Arbeit und Kultur. https://juliananslinger.at/wp-content/uploads/2023/02/Electronic-Working-Paper_Anslinger_FaireKI-wie-geht-das.pdf

Antunes, H. S., Freitas, P. M., Oliveira, A. L., Pereira, C. M., Sequeira, E. V., & Xavier, L. B. (Hrsg.). (2024). Multidisciplinary perspectives on artificial intelligence and the law (Law, Governance and Technology Series, Bd. 58). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-41264-6

Aysolmaz, B., Dau, N., & Iren, D. (2020). Preventing algorithmic bias in the development of algorithmic decision-making systems: A Delphi study. Proceedings of the 53rd Hawaii International Conference on System Sciences, 5267–5276. https://hdl.handle.net/10125/64390

Busse, B., Kleiber, I., Eickhoff, F. C., & Andree, K. (2023). Hinweise zu textgenerierenden KI-Systemen im Kontext von Lehre und Lernen. University of Cologne. https://zfl-lernen.de/wp-content/uploads/Uni_Koeln_Prorektorat_2023-02-02-Papier-Textgenerierende-KI-Systeme-Lehre-Lernen-1.pdf

Danks, D., & London, A. J. (2017). Algorithmic bias in autonomous systems. In International Joint Conference on Artificial Intelligence (26. Aufl., S. 1–7). IJCAI. https://www.cmu.edu/dietrich/philosophy/docs/london/IJCAI17-AlgorithmicBias-Distrib.pdf

Ferrara, E. (2023). Fairness and bias in artificial intelligence: A brief survey of sources, impacts, and mitigation strategies. Thomas Lord Department of Computer Science, USC Viterbi School of Engineering, University of Southern California. https://arxiv.org/pdf/2304.07683

Heesen, J., Bieber, C., Grunwald, A., Matzner, T., & Roßnagel, A. (2021). KI-Systeme und die individuelle Wahlentscheidung. Lernende Systeme. https://www.plattform-lernende-systeme.de/files/Downloads/Publikationen/AG3_WP_KI_und_Wahlen.pdf

Krüger, J., & Lischka, K. (2018). Damit Maschinen den Menschen dienen: Lösungsansätze, um algorithmische Entscheidungen in den Dienst der Gesellschaft zu stellen (Arbeitspapier Nr. 6). Bertelsmann Stiftung. https://doi.org/10.11586/2018019

Lischka, K., & Stöcker, P. D. C. (2017). Digitale Öffentlichkeit. Bertelsmann Stiftung. https://www.reframetech.de/wp-content/uploads/sites/23/2017/08/Digitale_Oeffentlichkeit_final.pdf

Melcher, P. R. (2025). KÜNSTLICHE INTELLIGENZ (KI) IM WISSENS- UND WISSENSCHAFTSMANAGEMENT [Dissertation, Hochschule Bonn-Rhein-Sieg, FB IWK]. Governance & Management. https://pub.h-brs.de/files/9216/wima_2025_melcher.pdf

Mohabbat Kar, R. (Hrsg.), Thapa, B. E. P. (Hrsg.), & Parycek, P. (Hrsg.). (2018). Algorithmische Entscheidungsfindung. Bundesverband der Verbraucherzentralen und Verbraucherverbände - Verbraucherzentrale Bundesverband e.V. https://www.ssoar.info/ssoar/bitstream/handle/document/57619/ssoar-2018-thapa-Thesenpapier_Algorithmische_Entscheidungsfindung.pdf?sequence=1

Müller, F., Schüßler, M., & Kirchner, E. (2020). Die Regulierung Künstlicher Intelligenz - Neuer Rechtsrahmen für Algorithmische Entscheidungssysteme? (Weizenbaum Series, 12). Weizenbaum Institute. https://doi.org/10.34669/wi.ws/12

Orwat, C. (2019). Diskriminierungsrisiken durch Verwendung von Algorithmen (1. Aufl.). Nomos. https://www.antidiskriminierungsstelle.de/SharedDocs/downloads/DE/publikationen/Expertisen/studie_diskriminierungsrisiken_durch_verwendung_von_algorithmen.pdf?__blob=publicationFile&v=3

Rathgeb, C. (2023). Diskriminierende KI: Ursachen und Lösungsansätze. Hochschule Darmstadt University of Applied Sciences, ATHENE National Research Center for Applied Cybersecurity. https://innen.hessen.de/x-myracloud-52e10ab1db13b0c6d9fd49fc11ce54e2/MzEzYTFhNGY2MTI4MjEwMGh0dHBzOi8vaW5uZW4uaGVzc2VuLmRlL3NpdGVzL2lubmVuLmhlc3NlbmRhdGVibl9wcm9mLl9kci5fY2hyaXN0aWFuX3JhdGhnZWJfLV8xNi4xMS4yMDIzLnBkZg==

Rebstadt, J. (2023). Trustworthy Artificial Intelligence Systems Engineering [Dissertation, Universität Osnabrück]. https://osnadocs.ub.uni-osnabrueck.de/bitstream/ds-202309019655/2/thesis_rebstadt.pdf

Rönnecke, S. (2024). Die europäische KI-Verordnung. Ein Weg hin zu

diskriminierungs-freien Algorithmen? Journal Netzwerk Frauen- und Geschlechterforschung NRW, Nr. 55/2024, 63-69. https://doi.org/10.17185/duepublico/82756

Smith, G., & Rustagi, I. (2020). Mitigating Bias in Artificial Intelligence: An Equity Fluent Leadership Playbook. Berkeley Haas Center for Equity, Gender and Leadership. https://haas.berkeley.edu/wp-content/uploads/UCB_Playbook_R10_V2_spreads2.pdf

Yapo, A., & Weiss, J. (2018). Ethical implications of bias in machine learning. Proceedings of the 51st Hawaii International Conference on System Sciences, 5365–5372. HICSS. https://scholarspace.manoa.hawaii.edu/bitstream/10125/50557/1/paper0670.pdf

Plagiatserklärung

Ich versichere, dass ich diese Arbeit selbständig angefertigt und keine anderen als die angegebenen Quellen benutzt habe.

Alle Stellen, die dem Wortlaut oder dem Sinn nach anderen Werken entnommen sind, habe ich in jedem einzelnen Fall unter genauer Angabe der Quelle (einschließlich des World Wide Web sowie anderer elektronischer Datensammlungen) deutlich als Entlehnung kenntlich gemacht. Dies gilt auch für angefügte Zeichnungen, bildliche Darstellungen, Skizzen und dergleichen.

Die vorliegende Arbeit wurde hinsichtlich Titel, Fragestellung, Aufbau und Inhalt, oder in umfangreichen Teilen und Auszügen daraus, noch nicht in einem Studiengang an dieser, oder einer anderen Hochschule, zur Anrechnung von Leistungspunkten vorgelegt.

Ich nehme zur Kenntnis, dass die nachgewiesene Unterlassung der Herkunftsangabe als versuchte Täuschung bzw. als Plagiat gewertet wird.

XXXX, den XX.XX.XXX