



Hochschule für Angewandte Wissenschaften Hamburg  
*Hamburg University of Applied Sciences*

## Bachelorarbeit

Kevin Kara

# Die Auswirkungen selbstlernender algorithmischer Entscheidungssysteme und wie man diese mindern kann

*Fakultät Technik und Informatik  
Department Informatik*

*Faculty of Engineering and Computer Science  
Department Computer Science*

Kevin Kara

Die Auswirkungen selbstlernender algorithmischer Entscheidungssysteme und wie man diese mindern kann

Bachelorarbeit eingereicht im Rahmen der Bachelorprüfung  
im Studiengang *Bachelor of Science Wirtschaftsinformatik*  
am Department Informatik  
der Fakultät Technik und Informatik  
der Hochschule für Angewandte Wissenschaften Hamburg

Betreuender Prüfer: Prof. Dr. Stefan Sarstedt

Zweitgutachter: Prof. Dr. Michael Neitzke

Eingereicht am: 03.08.2023

**Kevin Kara**

**Thema der Arbeit**

Die Auswirkungen selbstlernender algorithmischer Entscheidungssysteme und wie man diese mindern kann

**Stichworte**

Künstliche Intelligenz, maschinelles Lernen, menschliche Voreingenommenheit, algorithmische Voreingenommenheit, selbstlernende algorithmische Entscheidungssysteme

**Kurzzusammenfassung**

Selbstlernende algorithmische Entscheidungssysteme treffen Entscheidungen über Menschen, welche wesentliche Auswirkungen auf deren Leben haben. Aufgrund von algorithmischer Voreingenommenheit können diese Entscheidungen bestimmte Personengruppen diskriminieren. In dieser Arbeit werden Fallbeispiele der verheerenden Auswirkungen solcher voreingenommenen Entscheidungssysteme vorgestellt. Folglich liegt der Fokus auf den menschlichen, technischen und rechtlichen Ursachen sowie Lösungen dieser Problemstellung.

**Kevin Kara**

**Title of Thesis**

The effects of self-learning algorithmic decision systems and how to mitigate them

**Keywords**

Artificial intelligence, machine learning, human bias, algorithmic bias, self-learning algorithmic decision systems

**Abstract**

Self-learning algorithmic decision systems make decisions about people, which have a significant impact on their lives. Due to algorithmic bias, these decisions can discriminate against certain groups of people. In this paper, case studies with devastating effects of such biased decision systems will be presented. Consequently, the focus is on the human, technical and legal causes as well as solutions to this problem.

# Inhaltsverzeichnis

<b>Abbildungsverzeichnis</b> .....	<b>viii</b>
<b>Tabellenverzeichnis</b> .....	<b>x</b>
<b>Abkürzungsverzeichnis</b> .....	<b>xi</b>
<b>1 Einleitung</b> .....	<b>1</b>
1.1 Ziel dieser Arbeit.....	2
1.2 Aufbau dieser Arbeit.....	2
<b>2 Einführung in die Künstliche Intelligenz</b> .....	<b>5</b>
2.1 Künstliche Intelligenz (KI).....	5
2.1.1 KI-Arten.....	5
2.1.2 KI-Typen.....	6
2.2 Historie der KI.....	7
<b>3 Begrifflichkeiten</b> .....	<b>11</b>
3.1 Maschinelles Lernen.....	11
3.2 Neuronale Netze.....	14
3.3 Selbstlernende algorithmische Entscheidungssysteme.....	16
3.4 Bewusste- sowie unbewusste Voreingenommenheit.....	17
3.5 Algorithmische Voreingenommenheit.....	19
3.6 Stereotypisierung.....	19
3.7 Diskriminierung.....	20
3.7.1 Direkte Diskriminierung.....	22
3.7.2 Indirekte Diskriminierung.....	22
3.8 Fairness.....	23
<b>4 Vorteile von KI-Systemen bzw. AES</b> .....	<b>24</b>
<b>5 Ursachen kognitiver Verzerrung</b> .....	<b>26</b>

5.1	Übermaß an Informationen .....	27
5.1.1	Informationen, die mit bestehenden Informationen zusammenhängen.....	27
5.1.2	Informationen, welche bestehende Informationen aktualisieren.....	27
5.1.3	Informationen, welche bestehende Meinung bestätigen bzw. verstärken .....	28
5.1.4	Menschen erkennen eigene Denkfehler nicht .....	29
5.2	Nicht genügend Bedeutung .....	29
5.2.1	Informationen, welche als Lückenfüller fungieren .....	30
5.2.2	Informationen, die zu einer Person fehlen, würden hinzudedacht werden.....	30
5.3	Die Notwendigkeit, schnell zu handeln.....	31
5.3.1	Die Eigenschaft, zu handeln.....	31
5.3.2	Die Eigenschaft, fokussiert zu bleiben .....	32
5.3.3	Die Eigenschaft, Fehler vermeiden zu wollen.....	32
5.4	Verallgemeinerungen werden eher gespeichert als Details.....	32
5.4.1	Die Eigenschaft der Verallgemeinerung .....	33
5.4.2	Die Eigenschaft, Erinnerungen zu verstärken .....	33
<b>6</b>	<b>Voreingenommenheiten bezüglich der geschützten Merkmale eines Menschen.....</b>	<b>35</b>
6.1	Rassistische Voreingenommenheit .....	35
6.2	Geschlechterspezifische Voreingenommenheit .....	36
6.3	Altersvoreingenommenheit .....	37
6.4	Die Voreingenommenheit bezüglich sexueller Orientierung .....	39
6.5	Die Voreingenommenheit gegenüber Behinderten .....	40
<b>7</b>	<b>Sektoren mit voreingenommener Künstlicher Intelligenz.....</b>	<b>43</b>
7.1	Beschäftigungssektor .....	43
7.2	Finanzsektor .....	44
7.3	Gesundheitswesen .....	45
7.4	Rechtswesen.....	46
<b>8</b>	<b>Fallbeispiele .....</b>	<b>49</b>
8.1	AMAS .....	49
8.1.1	Wie funktioniert das System? .....	51
8.1.2	Welche Diskriminierungen liegen vor?.....	53
8.2	COMPAS .....	57

8.2.1	Wie funktioniert das System? .....	58
8.2.2	Welche Diskriminierungen liegen vor?.....	59
8.3	Amazons KI-Rekrutierungsunterstützungstool .....	63
8.3.1	Wie funktioniert das System? .....	64
8.3.2	Welche Diskriminierungen liegen vor?.....	64
8.4	Chat-GPT .....	65
8.4.1	Wie funktioniert das System? .....	67
8.4.2	Welche Diskriminierungen liegen vor?.....	69
<b>9</b>	<b>Ursachen algorithmischer Voreingenommenheit.....</b>	<b>76</b>
9.1	Entstehung eines KI-Modells .....	76
9.2	Menschliche Voreingenommenheit.....	78
9.3	Daten .....	78
9.3.1	Historische Voreingenommenheit.....	79
9.3.2	Feedback-Schleife .....	80
9.3.3	Korrelation und Kausalität .....	83
9.3.4	Proxy-Variablen .....	84
9.3.5	Über- und Unterrepräsentierung.....	86
9.4	KI-Entwickler und Situation der KI-Branche .....	88
9.4.1	Datenwissenschaftler (Data Scientist).....	88
9.5	Komplexität und Transparenz .....	97
9.5.1	KI-Modelle als Black-Box .....	98
9.5.2	Intransparenz von KI-Modellen .....	99
9.6	Benutzerfehler .....	100
9.6.1	Übertragung von Kontextverzerrungen.....	101
9.6.2	Verzerrung durch Missinterpretation .....	102
9.7	Rechtliche Regelungen.....	104
<b>10</b>	<b>Identifizierung &amp; Minderung von algorithmischer Voreingenommenheit.....</b>	<b>112</b>
10.1	Menschliche Maßnahmen .....	113
10.1.1	Diversität .....	113
10.1.2	Bewusstsein.....	115
10.1.3	Training zur Bekämpfung kognitiver Verzerrungen .....	118

10.1.4	Expertise.....	119
10.2	Datentechnische Maßnahmen .....	121
10.2.1	Vielfältige sowie ausgewogene Datensätze .....	121
10.2.2	Einbeziehung geschützter Merkmale .....	123
10.3	Statistische Metriken zur Messung der Genauigkeit & Fairness .....	125
10.3.1	Confusion-Matrix & Allgemeine statistische Metriken .....	126
10.3.2	Fairness-Metriken.....	130
10.4	Weitere datentechnische Maßnahmen.....	134
10.4.1	Pre-Processing.....	135
10.4.2	In-Processing.....	139
10.4.3	Post-Processing .....	141
10.5	Nachvollziehbarkeit trotz Intransparenz und Komplexität .....	143
10.6	Unterstützende Tools .....	145
10.7	Überwachen & Überprüfen des KI-Systems.....	147
10.8	Rechtliche Maßnahmen.....	149
10.8.1	Transparenz .....	150
10.8.2	Bekämpfung der Ohnmacht von Betroffenen .....	152
10.8.3	Kontrolle und Regulierung.....	153
<b>11</b>	<b>Fazit &amp; Ausblick.....</b>	<b>156</b>
11.1	Fazit.....	156
11.2	Ausblick .....	160
	<b>Literaturverzeichnis.....</b>	<b>162</b>

# Abbildungsverzeichnis

Abbildung 1: Historie der KI (eigene Darstellung).....	10
Abbildung 2: überwachtes Lernen (eigene Darstellung) (Wuttke, 2023a).....	13
Abbildung 3: unüberwachtes Lernen (eigene Darstellung) (Wuttke, 2023a).....	14
Abbildung 4: Neuronales Netz (eigene Darstellung) (Wuttke, 2023b).....	16
Abbildung 5: Confirmation Bias (eigene Darstellung) (Confirmation Bias: Warum der Bestätigungsfehler so gefährlich ist, o. D.) .....	29
Abbildung 6: Merkmale und Ausprägungen des AMAS (eigene Darstellung) (Allhutter, 2021) .....	52
Abbildung 7: Warnung Chat-GPT ( <i>ChatGPT</i> , 2023) .....	69
Abbildung 8: Schutzmechanismus Chat-GPT ( <i>ChatGPT</i> , 2023).....	70
Abbildung 9: Bevorzugung hellhäutiger Wissenschaftler ( <i>Is ChatGPT Woke And Biased? 16 examples that prove so</i> , 2023).....	71
Abbildung 10: Gedicht über Trump vs. Biden (Wolf, 2023) .....	72
Abbildung 11: Entwicklungsprozess KI-System (eigene Darstellung) (Machill, 2020; Stadler, 2020) .....	76
Abbildung 12: Feedback-Schleife (eigene Darstellung) (Marques, 2021).....	81
Abbildung 13: Geschützte Merkmale laut Deutschem Gesetz (eigene Darstellung) (Kolleck & Orwat, 2020) .....	106
Abbildung 14: Confusion Matrix (eigene Darstellung) (Narkhede, 2018) .....	127
Abbildung 15: Meistzitierten Fairness-Metriken (eigene Darstellung) (Machill, 2020; Verma & Rubin, 2018).....	130



Abbildung 16: Schritte der Datenvorverarbeitung (eigene Darstellung) (Lawton, 2023;  
Mesevage, 2021) ..... 135

# Tabellenverzeichnis

Tabelle 1: Allgemeine statistische Metriken (Narkhede, 2018; Oppermann, 2021; Verma & Rubin, 2018).....	130
---	-----

# Abkürzungsverzeichnis

KI	Künstliche Intelligenz
ML	machine learning / maschinelles Lernen
AES	selbstlernende algorithmische Entscheidungssysteme
NN	neuronale Netze

# 1 Einleitung

Künstliche Intelligenz ist, aufgrund der Durchbrüche in jüngster Vergangenheit und ihrer vielfältigen Nutzungsmöglichkeiten, eine hochaktuelle Thematik, welche viele Menschen weltweit betrifft. Es existieren etliche Beispiele wie künstliche Intelligenz das Leben eines Menschen beeinflusst. Das zurzeit wohl bekannteste Beispiel ist das KI-Tool Chat-GPT. Dieses wird täglich von Millionen von Nutzern verwendet, um sie unter anderem im Alltag oder auf der Arbeit zu assistieren (Clickblogger, 2022; Linden, 2023; Fam, 2023).

Innerhalb des Themengebietes der künstlichen Intelligenz fokussiert sich diese Arbeit jedoch auf spezielle selbstlernende Systeme, welche Entscheidungen über Menschen treffen, die das Leben dieser stark beeinflusst. Beispielsweise wird solch ein System bei der Verurteilung von Straftätern zu Hilfe gezogen (Rütten, 2018; Öktem, 2022).

Diese Systeme werden selbstlernende algorithmische Entscheidungssysteme genannt und unter anderem aus dem Grund verwendet, für neutralere Entscheidungen zu sorgen. Allerdings liegen, aufgrund der Tatsache, dass Menschen meist vorurteilsbehaftet sind und diese Voreingenommenheiten in das Verhalten und folglich in die Entscheidungsfindung dieser Systeme einfließen kann, mit diesen Entscheidungssystemen oftmals keine neutralen Entscheider vor, sondern vielmehr verzerrte Maschinen, die bestehende Vorurteile der Menschen widerspiegeln und verstärken. Dies wird auch als algorithmische Voreingenommenheit bezeichnet. Dementsprechend stehen in dieser Arbeit algorithmisch voreingenommene selbstlernende Entscheidungssysteme, die Auswirkungen durch ihren Einsatz und Wege zur Minderung dieser Voreingenommenheiten im Fokus (Rosso, 2018; Marques, 2021; Öktem, 2022; Kilian, 2022; Bär, 2022).

## **1.1 Ziel dieser Arbeit**

Die bestehenden wissenschaftlichen Artikel zu diesem Thema konnten nie einen absoluten Lösungsvorschlag vorlegen, da die Thematik nicht umfassend genug erläutert wurde und somit in den Lösungen nicht jeder relevante Aspekt berücksichtigt wurde. Aufgrund der Tatsache, dass algorithmische Voreingenommenheit über diverse Wege in diese Entscheidungssysteme gelangen kann, ist ein breit gefächertes Lösungsvorschlag von Nöten, um dieser Thematik wirkungsvoll entgegenwirken zu können (De Montis, 2019; Machill, 2020).

Aus diesem Grund ist das Ziel dieser Arbeit, die bestehenden wissenschaftlichen Beiträge zu diesem Thema zusammenzuführen und zu vervollständigen, um so eine wissenschaftliche Arbeit hervorzubringen, die das Thema der algorithmischen Voreingenommenheit von selbstlernenden Entscheidungssystemen allumfassend erfasst und einen vielfältigen Lösungsvorschlag vorlegt. Diese Arbeit soll insbesondere verdeutlichen, wie die Vorurteile der Menschen in algorithmischer Voreingenommenheit resultieren, welche teils verheerenden Auswirkungen diese voreingenommenen Entscheidungssysteme auf Menschen haben können und was die menschlichen, technischen und rechtlichen Ursachen sowie Lösungen für diese Thematik sind (Bär, 2022; Machill, 2020; Marques, 2021; Rosso, 2018).

Das Ergebnis dieser Arbeit soll somit als Orientierungshilfe und Grundlage für jegliche Personen dienen, die sich mit der algorithmischen Voreingenommenheit von selbstlernenden Entscheidungssystemen auseinandersetzen.

## **1.2 Aufbau dieser Arbeit**

Im Folgenden werden kurz die zu erwartenden Inhalte der jeweiligen Kapitel nähergebracht.

Angefangen mit diesem Kapitel, der Einleitung. Dieses Kapitel fungiert einführend in das hochaktuelle Forschungsfeld und die Menschenleben relevante Thematik dieser Arbeit. Weiterführend hebt es einerseits die Unvollständigkeit vorhandener wissenschaftlicher Beiträge und andererseits die Notwendigkeit einer allumfassenderen wissenschaftlichen Arbeit, welche mit dieser erreicht werden soll, hervor.

Als nächstes folgt eine Einführung in die künstliche Intelligenz, in welcher die Begrifflichkeit sowie die verschiedenen Arten sowie Typen dessen näher erläutert werden. Weiterführend wird in diesem Kapitel die Historie der künstlichen Intelligenz veranschaulicht. Somit wird das, für diese Arbeit relevante, Forschungsfeld ausgiebig vorgestellt (Anyoha, 2017; Burns, Laskowski & Tucci, 2023; Schick, 2018).

Daraufhin folgt ein Kapitel, in welchem die für diese Arbeit fundamental relevanten Begrifflichkeiten vorgestellt und erklärt werden. Um den Umfang dieses Kapitels nicht zu sprengen, wurden Begriffe, die in einem speziellen Kontext Verwendung finden, erst an der jeweiligen Stelle des spezifischen Kontextes vorgestellt, da sie vorher nicht benötigt werden.

Anschließend folgt das Kapitel „Vorteile von KI-Systemen“. Dieses veranschaulicht die Nützlichkeit und Stärken von Systemen der künstlichen Intelligenz bzw. algorithmischen Entscheidungssystemen (Clickblogger, 2022; Kilian, 2022).

Im anschließenden Kapitel werden die Ursachen für kognitive Voreingenommenheiten präsentiert, wobei insbesondere deren Unvermeidbarkeit hervorgehoben wird. Weiterführend werden unterschiedlichste kognitive Verzerrungen vorgestellt, die teilweise im späteren Verlauf der Arbeit nochmals aufgegriffen werden (Bär, 2022; Benson, 2016; Rosso, 2018).

Nachfolgend werden die Vorurteile bezüglich der geschützten Merkmale eines Menschen vorgestellt, welche aus den kognitiven Voreingenommenheiten der Menschen resultieren und welche in Sektoren wie beispielsweise dem Beschäftigungssektor aufkommen (Benson, 2016; Machill, 2020).

Daraufhin werden die Sektoren vorgestellt, in denen voreingenommene algorithmische Entscheidungssysteme besonders Anwendung finden, da dort vor Nutzung dieser Systeme bereits Voreingenommenheiten vorherrschen (Kolleck & Orwat, 2020; Machill, 2020).

Im nächsten Kapitel werden einige Fallbeispiele präsentiert, die teilweise innerhalb der angesprochenen Sektoren Verwendung finden oder fanden. Mit diesen sollen die verheerenden Auswirkungen von algorithmischer Voreingenommenheit in selbstlernenden Entscheidungssystemen illustriert werden (Beck, 2023; De Montis, 2019; Kolleck & Orwat, 2020).

Folglich werden im nächsten Kapitel die diversen Ursachen vorgestellt, die veranschaulichen, wie algorithmische Voreingenommenheiten in selbstlernende algorithmische Entscheidungssysteme gelangen können. Der Fokus liegt hierbei darauf wie menschliche voreingenommene Handlungen, unausgewogene Datensätze und das Lernverhalten des Systems in algorithmischen Verzerrungen münden kann. Darüber hinaus wird aufgezeigt, wie sich diese Verzerrungen im KI-System verfestigen können (De Montis, 2019; Machill, 2020; Marques, 2021; Mohammad, 2021).

Im vorletzten Kapitel werden dann Lösungsmöglichkeiten und -vorschläge präsentiert, mit dem Ziel aufzuzeigen, wie algorithmische Voreingenommenheit gemindert werden kann. Hierbei liegt der Fokus einerseits auf Möglichkeiten der Ermittlung und Minderung von menschlichen Voreingenommenheiten und andererseits auf datentechnischen Maßnahmen, die teilweise auf statistischen Metriken beruhen (De Montis, 2019; Fiegele, 2021; Gasser, 2019; Lawton, 2023; Machill, 2020; Marques, 2021; Mesevage, 2021; Mohammad, 2021; Verma & Rubin, 2018).

Außerdem wird die Relevanz von Überwachungs- sowie Überprüfungsmaßnahmen hervorgehoben. Weiterführend werden unterstützende Tools präsentiert, die bei der Bekämpfung von algorithmischer Voreingenommenheit helfen sollen, und zum Abschluss dieses Kapitels werden mögliche durchführbare rechtliche Maßnahmen vorgestellt (De Montis, 2019; Deoras, 2020; Öktem, 2022; Gasser, 2019; Kolleck & Orwat, 2020; Marques, 2021).

Im letzten Kapitel folgen sowohl das Fazit als auch der Ausblick. Hier werden die Kerninhalte der Arbeit zusammengefasst und es wird untersucht, ob das Ziel dieser Arbeit erreicht wurde. Weiterführend wird darauf eingegangen, wie Forschende dieses Gebietes diese Arbeit weiterbehandeln können.

Abschließend wäre darauf hinzuweisen, dass zum einen an einigen Stellen auf eine tiefgehende Erläuterung verzichtet wird, um den Rahmen dieser Arbeit nicht zu sprengen und zum anderen werden relevante Fachbegriffe und Namen beim erstmaligen Auftreten *kursiv* dargestellt, um diese hervorzuheben.

## 2 Einführung in die Künstliche Intelligenz

### 2.1 Künstliche Intelligenz (KI)

*Künstliche Intelligenz* oder abgekürzt *KI* ist ein Oberbegriff für Maschinen, welche das Ziel verfolgen, menschliche Intelligenz zu simulieren und somit Resultate hervorzubringen, die auch ein Mensch hätte hervorbringen können. Hierfür ist KI in der Lage, aus Daten sowie früheren Erlebnissen zu lernen, mithilfe von Regeln zu Schlussfolgerungen zu gelangen und diese zu evaluieren, um sich selbst weiterzuentwickeln und zu verbessern (Gasser, 2019; Schick, 2018).

Dementsprechend unterscheiden sich KI getriebene Maschinen von Maschinen, bei welchen der Output zu einem gegebenen Input immer derselbe ist, da diese nicht eigenständig lernen, sondern jegliche ihrer Handlungen fest vorprogrammiert sind (Gasser, 2019; Schick, 2018).

#### 2.1.1 KI-Arten

KI kann in zwei Arten kategorisiert werden. Die eine Kategorie wird *narrow AI* oder auch *schwache KI* genannt. Diese Art von KI wird dahingehend trainiert, eine spezifische Aufgabe zu erfüllen. Jedoch kann diese nicht auf Problemstellungen außerhalb des fokussierten Bereiches angewandt werden, da sie ihre Methodik der Problemlösung nicht auf andere Bereiche übertragen kann (Gasser, 2019; Mareike, 2019).

Ein Beispiel für solch eine Art von KI wäre das Schach spielende Computerprogramm namens *Deep Blue*. Dieses ist darauf spezialisiert, Schachpartien zu gewinnen und somit stets den bestmöglichen Zug durchzuführen. Nichtsdestotrotz ist diese KI und sein Lernalgorithmus speziell auf Schach ausgelegt, sodass die KI sich mit ihrer erlernten Herangehensweise auf individuelle



Problemstellungen innerhalb einer Schachpartie fokussiert. Außerhalb des Spektrums von Schach kann diese KI allerdings nicht ihre Methodik auf andere Bereiche anwenden (Burns et al., 2023; Gasser, 2019; Mareike, 2019).

Die andere Kategorie wird *general AI* oder *starke KI* bezeichnet. Eine KI dieser Art wäre in der Lage, Problemstellungen jeglicher Ausprägung anzugehen, da sie die intellektuellen Fähigkeiten des Menschen repliziert oder sogar erweitert. Bis zum heutigen Zeitpunkt konnte keine solche KI erschaffen werden (Burns et al., 2023; Gasser, 2019; Mareike, 2019).

### **2.1.2 KI-Typen**

Weiterführend kann künstliche Intelligenz in vier Typen eingeordnet werden (Burns et al., 2023; Hintze, 2016).

#### **Typ 1 – Reaktive Maschinen**

Diese KI fokussiert sich auf eine spezifische Aufgabe und kann wie z. B. das Deep Blue Programm Schachfiguren und ihre Fähigkeiten identifizieren und auf dieser Grundlage den nächsten Schachzug prognostizieren. Nichtsdestotrotz reagiert dieses Programm lediglich auf das Verhalten seines Gegners, denn es kann keine Erfahrung aus früheren Partien heranziehen, da es kein Gedächtnis besitzt (Burns et al., 2023; Hintze, 2016; Mareike, 2019).

#### **Typ 2 – Begrenzter Speicher**

Diese KI bezieht Erfahrungswerte aus früheren Entscheidungen mit in die aktuelle Entscheidungsfindung ein. Demzufolge besitzt solch eine KI ein Gedächtnis und somit die Fähigkeit, gesammelte Erfahrung aus vergangenen Handlungen bei Bedarf in die aktuelle Lösungsfindung einzubeziehen. Autonom fahrende Autos lassen sich in diesen Typ kategorisieren (Burns et al., 2023; Hintze, 2016; Mareike, 2019).

### **Typ 3 – Theorie des Verstandes**

KI dieses Typs ist in der Lage, menschliche Emotionen zu verstehen und zu erkennen. Unter Berücksichtigung der Tatsache, dass Menschen ihr Verhalten von ihren Emotionen und aktuellem Gefühlszustand beeinflussen lassen, kann solch eine KI anhand des emotionalen Zustandes eines Menschen vorhersagen, welche Handlungen dieser wahrscheinlich als nächstes vollziehen wird (Burns et al., 2023; Hintze, 2016; Mareike, 2019).

Dieser Typ existiert zum jetzigen Zeitpunkt noch nicht, jedoch wäre die Umsetzung dieses Typs der nächste Schritt für die Vision einer Welt, in welcher KI getriebene Maschinen mit Menschen kommunizieren und zusammenarbeiten (Burns et al., 2023; Hintze, 2016; Mareike, 2019).

### **Typ 4 - Selbstwahrnehmung**

Dieser Typ von KI ist so weit entwickelt, dass er sich selbst wahrnehmen kann. Dieser weist also ein Bewusstsein für sich sowie seine Umwelt auf und kann somit seinen eigenen Zustand als auch den Zustand anderer wahrnehmen. Solch eine KI existiert ebenfalls noch nicht, da sie auf Typ 3 aufbaut (Burns et al., 2023; Hintze, 2016; Mareike, 2019).

## **2.2 Historie der KI**

Bereits in der ersten Hälfte des 20. Jahrhunderts wurden autonom denkende und intelligente Roboter in Filmen wie *Der Zauberer von Oz* oder *Metropolis* illustriert. Dies ließ Forscher, darunter auch *Alan Turing*, einen der einflussreichsten Informatiker und Theoretiker, ins Grübeln bringen. Laut Turing nutzen Menschen zugreifbare Informationen sowie den Verstand, um Probleme zu lösen und Entscheidungen zu treffen (Anyoha, 2017; Burns et al., 2023; Machill, 2020).

Auf dieser Grundlage erstellte er in den 1950ern einen wissenschaftlichen Artikel namens *Computing Machinery and Intelligence*, in welchem er erörtert, wie intelligente Maschinen

erzeugt und getestet werden können. Seine in diesem Artikel niedergelegte Idee spricht von einer universellen Maschine, die alle Berechnungen durchführen könnte, die auch von Menschen durchgeführt werden können. Sein Ziel war es, die menschliche Methodik der Problemlösung auf Maschinen zu übertragen. Aufgrund der damals teuren Rechner, welche zudem geringe Rechenleistung aufwiesen, konnte Turing seinem Konzept nicht nachgehen (Anyoha, 2017; Burns et al., 2023; Machill, 2020).

Einige Jahre später, im Jahr 1956, wurde in der *Dartmouth Summer Research Project on Artificial Intelligence-Konferenz* das weltweit erste KI-Programm namens *Logic Theorist* vorgestellt. Dieses Programm ahmt die menschliche Methodik der Problemlösung nach (Anyoha, 2017; Burns et al., 2023; Machill, 2020).

Diese Konferenz wird als die Geburtsstunde von künstlicher Intelligenz verstanden. Der damalige Gastgeber der Veranstaltung, *John McCarthy*, prägte den Begriff künstliche Intelligenz und wird somit als Gründungsvater von künstlicher Intelligenz angesehen. Diese Konferenz sorgte dafür, dass KI in den Folgejahren zu einem trendigen Forschungsfeld wurde (Anyoha, 2017; Burns et al., 2023; Machill, 2020).

Rechner wurden zudem günstiger, konnten mehr Informationen speichern und die Rechenleistung stieg an. Leitende Forscher waren so überzeugt von den Möglichkeiten künstlicher Intelligenz und den damaligen technischen Mitteln, dass Forscher wie *Marvin Minsky* bereits im Jahr 1970 Aussagen trafen wie „from three to eight years we will have a machine with the general intelligence of an average human being“ (Machill, 2020, S. 5). Der grundlegende Beweis für selbstdenkende Maschinen und demzufolge KI wurde zwar gesetzt, jedoch konnten zum damaligen Zeitpunkt nicht die erwarteten Ziele der KI-Disziplin erreicht werden. Trotz der Forschungslust und den ambitionierten Zielen der Forscher lag das größte Hindernis bei den immer noch zu schwachen Rechnern der damaligen Zeit (Anyoha, 2017; Burns et al., 2023; Machill, 2020).

Somit verlor das Forschungsfeld in den Jahren von 1970 bis 1980 an Interesse. In den Jahren von 1990 bis 2000 waren Rechner ausgeprägt genug, um viele Meilensteine der KI zu setzen, insbesondere ein Ereignis sorgte für viel Aufmerksamkeit. Im Jahr 1997 konnte IBMs Schach spielendes Deep Blue-Programm den damaligen Weltmeister in Schach, *Gary Kasparov*, besiegen. Dies war das erste Mal, dass Gary gegen ein Computer-Programm eine Schachpartie

verlor. Dieses Ereignis zeigte der Welt, dass sich KI als Entscheidungsentität eignet (Anyoha, 2017; Burns et al., 2023; Machill, 2020).

Weitere Meilensteine wie eine Spracherkennungssoftware für Windows oder eine Software, die menschliche Emotionen erkennt und wiedergibt, wurden erreicht. Es stellte sich heraus, dass die Theorien für KI bereits vor Jahrzehnten gelegt wurden, diese jedoch ihrer Zeit voraus waren und damals noch nicht die Rechnerleistung und Speicherkraft zur Verfügung stand wie zu modernerer Zeit (Anyoha, 2017; Burns et al., 2023; Machill, 2020).

Ab den 2000ern tritt das *Moore'sche Gesetz* besonders in Kraft. Dieses besagt, dass sich die Rechenleistung von Computern jährlich verdoppelt. In der Realität tritt diese Verdopplung alle 18 Monate ein. Es wird jedoch trotzdem ein exponentieller Anstieg der Rechenleistung erkennbar. Weiterführend herrscht, ausgelöst durch die Digitalisierung, ebenfalls ein exponentielles Wachstum des Datenvolums, sodass die heutige Zeit als datengesteuertes Zeitalter bezeichnet wird (Greaton, 2019; Raitner, 2018).

Diese Unmengen an Daten können nicht von Menschen ohne Hilfe von Computern verarbeitet werden. Unter Berücksichtigung all dieser Aspekte kristallisiert sich die Relevanz von Computern hervor, die den Menschen bei der Verarbeitung dieser Daten helfen (Greaton, 2019; Raitner, 2018).

Die Historie zeigt auf, dass die Leistung KI getriebener Maschinen von der verfügbaren Rechenleistung abhängig ist. Des Weiteren lernen diese KI-Maschinen unter anderem anhand der ihnen zur Verfügung stehenden Daten. Da die Rechenleistung sowie die zur Verfügung stehenden Daten exponentiell zunehmen, werden zur heutigen Zeit kontinuierlich neue Meilensteine der KI gesetzt. Hier wären beispielsweise autonom fahrende Kraftfahrzeuge oder Chatbots wie *Chat-GPT*, welcher für fast jede Frage eine Antwort liefern kann, besonders erwähnenswert. Die Tatsache, dass Chat-GPT detaillierte Antworten zu einer Vielfalt an Themen liefern kann und trotzdem nur als schwache KI sowie Typ 2-KI klassifiziert wird, illustriert, wie hochaktuell sowie wichtig dieses Technologiefeld ist, welches einen immer größer werdenden Teil unseres alltäglichen Lebens einnimmt, und was noch alles in der Zukunft möglich sein könnte (Anyoha, 2017; Burns et al., 2023; Cashman, 2023; Greaton, 2019; Machill, 2020; Mareike, 2019; Naumann, 2023; Raitner, 2018).

Die Relevanz von KI spiegelt sich auch in den verschiedenen Sektoren unserer Gesellschaft wider. Wie das Deep Blue Computerprogramm 1997 bereits zeigte, eignet sich KI dort besonders gut als Entscheidungsentität. Folglich finden KI-Programme im Gesundheitswesen, Wirtschaftswesen, Beschäftigungssektor, Finanzwesen, Rechtswesen usw. Anwendung, sodass KI beispielsweise bei Kreditanträgen mitwirkt und direkt oder indirekt darüber entscheidet, ob eine Person ihren angefragten Kredit erlangt oder nicht. Solche Entscheidungen teils oder komplett an selbstlernende Computerprogramme auszulagern, hat seine Vorteile, aber auch Nachteile. Genauer hierzu wird im weiteren Verlauf dieser Arbeit näher erläutert (Anyoha, 2017; Burns et al., 2023; Machill, 2020).

Folgender Zeitstrahl fasst die Historie der KI kompakt zusammen.

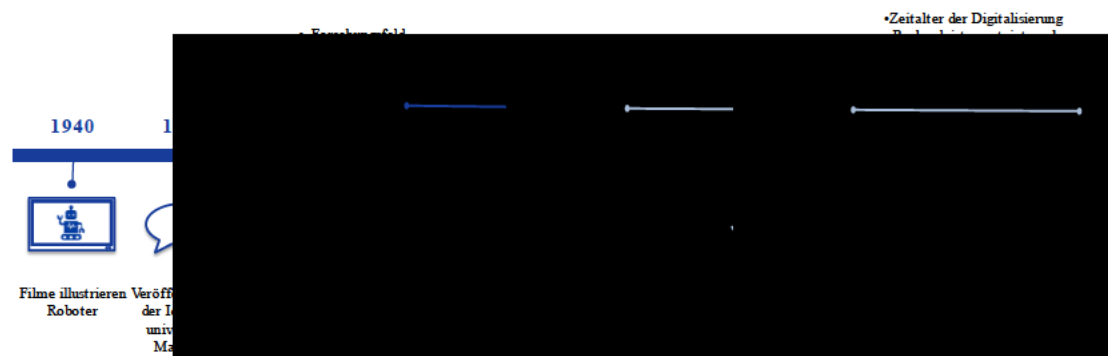


Abbildung 1: Historie der KI (eigene Darstellung)

## 3 Begrifflichkeiten

### 3.1 Maschinelles Lernen

Vorerst besteht jedoch Bedarf, auf weitere mit KI in Verbindung stehende Begriffe einzugehen: angefangen mit *maschinellern Lernen* oder *machine learning* abgekürzt *ML*. Dieses ist ein Teilgebiet der KI, welches den Maschinen die Möglichkeit bietet, anhand von Daten eigenständig zu lernen, ohne dass dieser Lernprozess ausdrücklich vorprogrammiert wurde. Mithilfe von maschinellem Lernen eignen sich Systeme die Methodik an, Muster sowie Zusammenhänge in den Daten zu erkennen, wodurch sie zu immer genaueren Ergebnissen kommen und sich somit verbessern. Ziel ist es, dass das System im Endeffekt anhand der gegebenen Inputdaten und der Problemstellung einen möglichst präzisen Output liefert (Gasser, 2019; Wuttke, 2023a).

Es existieren mehrere verschiedene Arten des maschinellen Lernens. Da sich diese Arbeit primär auf Programme fokussieren wird, welche Prognosen sowie Klassifikationen zum Ergebnis haben, konzentrieren wir uns im weiteren Verlauf auf die Art des *überwachten Lernens* (Wuttke, 2023a).

Anhand eines Beispiels kann dieses Konzept sowie die Idee des maschinellen Lernens nähergebracht werden. Angenommen, ein KI-Programm soll in der Lage sein, identifizieren zu können, ob es sich bei einem gegebenen Bild um einen Hund oder eine Katze handelt. Also müsste das Programm die Fähigkeit besitzen, die Eingabedaten (die Bilder) richtig in Hund oder Katze zu klassifizieren und dies als Output dem Nutzer mitzuteilen (Wuttke, 2022a). Nun würde man beim überwachten Lernen folgendermaßen vorgehen:

1. Zuerst einmal müssten zum Trainieren des Modells, also des KI-Programmes, *Trainingsdatensätze* herangezogen werden, bei welchen die Lösung, also ob es sich beim

gegebenen Input um einen Hund oder eine Katze handelt, bereits feststeht (Wuttke, 2022a).

2. Nun würde das Modell die Trainings-Bilder in Hund oder Katze klassifizieren. Anschließend wird geschaut, ob die Ergebnisse des Programmes mit den Lösungen der Trainingsdaten übereinstimmen. Anhand dessen lässt sich eine Aussage über die Genauigkeit des Programmes treffen. Diese ersten beiden Schritte sind beliebig häufig wiederholbar und in der Regel werden an dieser Stelle auch mehrere Iterationen durchgeführt, um die Genauigkeit des Programmes zu erhöhen (Wuttke, 2022a).

Im Prinzip sorgt das ständige Überprüfen der entschiedenen Klassifizierung des Programmes mit der tatsächlichen Lösung dazu, dass das KI-Programm bei Fehlentscheidung kontinuierlich mehr und neue Muster sowie Zusammenhänge sucht sowie erkennt, sodass er dementsprechend stetig seine Funktionen und Regeln anpasst. Auf diese Weise lernt und verbessert sich das KI-Programm eigenständig. Wichtig hervorzuheben wäre, dass solch eine KI-Maschine nur erzeugt werden kann, wenn der KI-Entwickler die Trainingsdaten zur Verfügung stellt (Wuttke, 2022a).

3. Ist der Entwickler zufrieden mit der Erfolgsrate des Programmes und zuversichtlich, dass es auf unbekannte neue Daten angewandt werden kann, so wird der Trainingsprozess beendet und das Modell wird mit neuen Inputdaten, nämlich den *Testdaten* getestet. Hier wird geprüft, wie genau die Ergebnisse des KI-Modells auf einem unbekanntem Datensatz sind (Wuttke, 2022a).
4. Sollte hier die Genauigkeit ebenfalls zufriedenstellend für den Entwickler sein, so kann man das KI-System fortan in der Praxis anwenden und seine Ergebnisse als Lösung für den jeweiligen Input nehmen (Wuttke, 2022a).

Die soeben beschriebene Lernmethode des überwachten Lernens kann mit folgender Abbildung veranschaulicht werden.

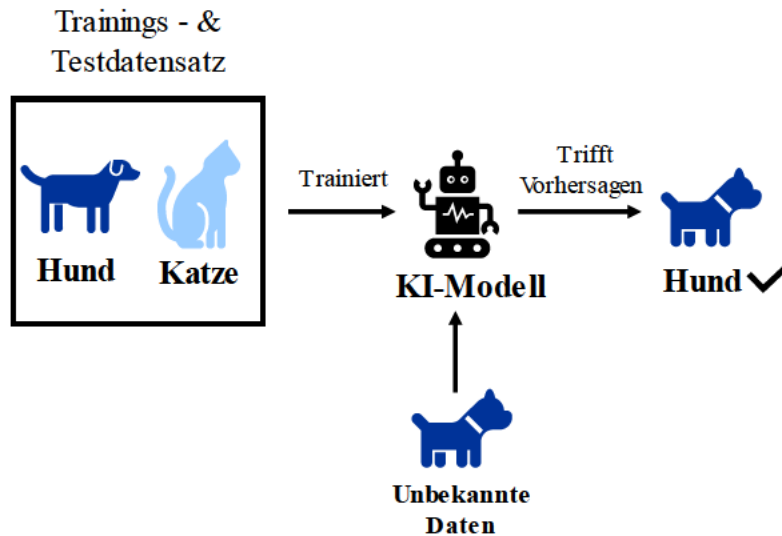


Abbildung 2: überwachtes Lernen (eigene Darstellung) (Wuttke, 2023a)

An dieser Stelle wäre der Unterschied zur *unüberwachten Lernmethode* hervorzuheben, da diese im Laufe dieser Arbeit ebenfalls Anwendung finden wird (Wuttke, 2023a).

Bei dieser Lernmethode erhält das KI-Modell keine speziell aufbereiteten Daten zum Erlernen eines diesen Daten möglichst ähnlichen Verhaltens. Sondern vielmehr würde das KI-Modell eine große Menge an Daten zur Verfügung gestellt bekommen mit dem Ziel, Muster sowie Zusammenhänge in diesen zu erkennen. Dementsprechend hätte das KI-Modell kein Endergebnis im Sinne einer Prognose vorliegen, wodurch es keine Möglichkeit gibt, wie beim überwachten Lernen die vom Modell erzeugten Ergebnisse mit der Zielvorstellung der Entwickler zu vergleichen. Aus diesem Grund wäre die Güte von Modellen, die auf dieser Lernmethode beruhen, schwieriger einzuschätzen als von Modellen, die auf dem überwachten Lernen beruhen (Wuttke, 2023a).



Die Lernmethode des unüberwachten Lernens kann mit folgender Abbildung veranschaulicht werden.

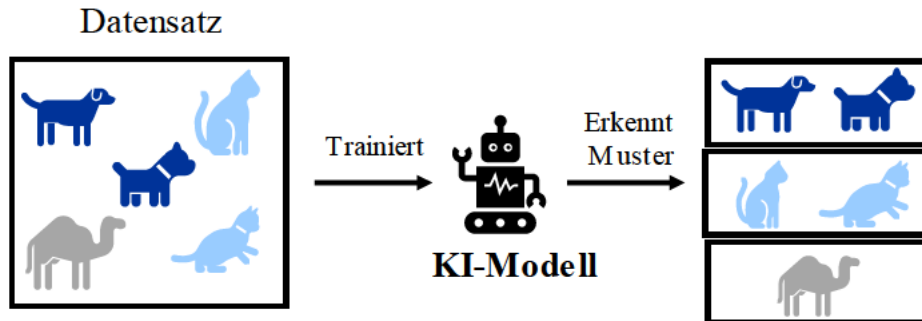


Abbildung 3: unüberwachtes Lernen (eigene Darstellung) (Wuttke, 2023a)

Die Erläuterung des überwachten- sowie unüberwachten Lernens und insbesondere das Beispiel der überwachenden Lernmethode, illustriert die Funktionsweise und Möglichkeiten des maschinellen Lernens.

Hiermit wurde nähergebracht, dass KI-Systeme abhängig von der jeweiligen Lernmethode ein bestimmtes Verhalten erlernen können. Nun wäre relevant zu wissen, wie das System sich das jeweilige Verhalten beibringt.

### 3.2 Neuronale Netze

An dieser Stelle wären somit *neuronale Netze* oder *NN* vorzustellen. NN fungieren ähnlich wie ein menschliches Gehirn und werden dafür eingesetzt komplexe Aufgaben zu lösen (Wuttke, 2023b).

Diese NN finden dementsprechend bei denjenigen KI-Systemen und zugrundeliegenden ML-Lernmethoden Anwendung, bei denen das System eine komplexe Aufgabe bewältigen soll. Zum Einordnen eines Tieres in Hund oder Katze kann ebenfalls ein NN verwendet werden (Wuttke, 2023b).

Es existieren verschiedene Arten von NN, welche unterschiedliche Vor- und Nachteile aufweisen, sodass je nach Aufgabenstellung ein geeignetes zu wählen wäre (Wuttke, 2023b).

Nichtsdestotrotz funktionieren diese NN im Kern alle ähnlich. Jedes NN weist *Neuronen* auf, welche Informationen annehmen, diese verarbeiten und weitergeben. Diese Neuronen sind miteinander verknüpft, sodass das vorangehende Neuron die Informationen für die nächsten Neuronen modifiziert. Hierbei existieren drei Schichten, in welche sich diese Neuronen einordnen lassen (Wuttke, 2023b).

Angefangen mit der *Eingabeschicht*. Hier finden sich diejenigen Neuronen wieder, die die Eingabeinformationen des NN annehmen, diese verarbeiten und an die nächste Schicht weitergeben (Wuttke, 2023b).

Diese ist die sogenannte *verborgene Schicht*. Hier empfangen die Neuronen die verarbeiteten Informationen der Eingabe-Neuronen und modifizieren diese weiter. Somit nehmen sie, genauso wie die Eingabe-Neuronen, auf den empfangenen Informationen Berechnungen sowie Gewichtungen vor und übergeben das Resultat an die nächsten Neuronen bzw. an die nächste Schicht weiter (Raveling, 2022; Wuttke, 2023b).

Die letzte Schicht ist die *Ausgabeschicht*. Die hier arbeitenden Neuronen enthalten das letztendliche Ergebnis des NN (Wuttke, 2023b).

Wichtig hervorzuheben ist, dass die Ein- sowie Ausgabeschicht nur eine Ebene an Neuronen aufweist, wohingegen die verborgene Schicht n-viele Ebenen an Neuronen aufweisen kann. Wobei  $n \geq 0$ . Aus diesem Grund wird diese Schicht auch so bezeichnet, da hier die hauptsächliche Komplexität solch eines NN liegt und es somit schwierig sein kann die konkrete Funktionsweise dieser Schicht nachzuvollziehen (Hunts, 2019; Wuttke, 2023b).

Diese soeben beschriebene Funktionsweise eines solchen NN wäre in Abbildung 4 illustriert.

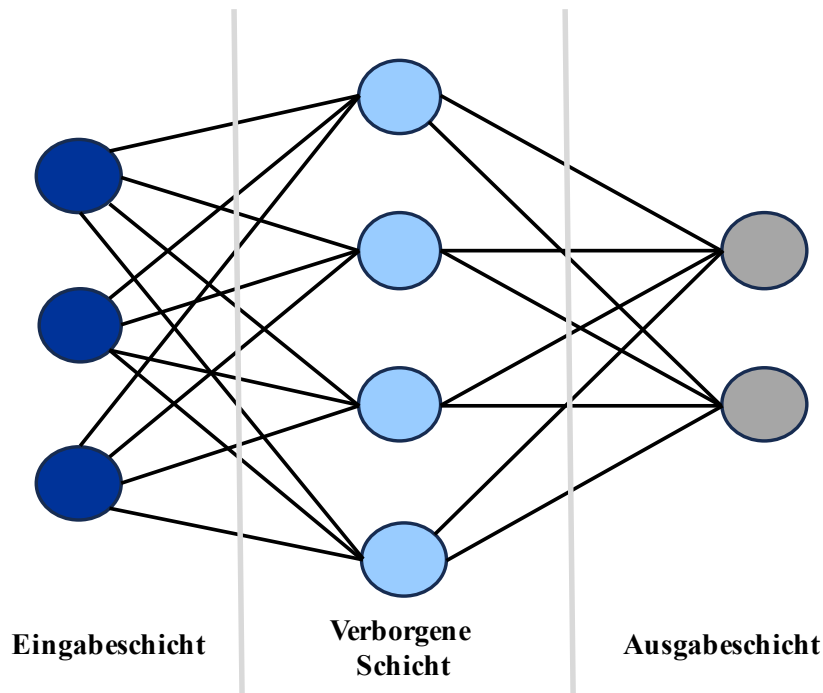


Abbildung 4: Neuronales Netz (eigene Darstellung) (Wuttke, 2023b)

Die mithilfe dieser neuronalen Netze erzeugten KI-Systeme können unter anderem als *selbstlernende Algorithmische Entscheidungssysteme* verstanden werden (Öktem, 2022; Litzel & Luber, 2019).

### 3.3 Selbstlernende algorithmische Entscheidungssysteme

*Algorithmische Entscheidungssysteme* sind im Kern Algorithmen, jedoch handelt es sich bei AES immer um Algorithmen, die dafür ausgelegt sind, eine Entscheidung oder Prognose zu treffen z. B. über die Frage: Sollte Person x einen Kredit erhalten oder nicht? Diese Entscheidungen oder Prognosen, z. B. Ja, Person x sollte einen Kredit erhalten, sind dann endgültig oder gelten zumindest als Handlungsempfehlung für Menschen, welche dann die letztendlich endgültige Entscheidung treffen (Öktem, 2022; Litzel & Luber, 2019; Machill, 2020).

Man spricht von selbstlernenden algorithmischen Entscheidungssystemen, wenn diese mithilfe von ML, insbesondere der überwachten Lernmethode, trainiert worden sind. Im weiteren Verlauf dieser Arbeit wird primär die Abkürzung *AES* für diese Systeme verwendet (Öktem, 2022; Litzel & Luber, 2019; Machill, 2020).

Somit sind mit AES, selbstlernende KI-Algorithmen gemeint, die zur Entscheidungsfindung personenbezogene Daten heranziehen und Entscheidungen oder Handlungsempfehlungen für Problemstellungen niederlegen, die direkte Auswirkungen auf Menschenleben haben. Somit liegen mit AES, KI-Systeme vor, die für Menschen und ihre Rechte wichtig sind, da sie menschenrechtlich relevante Entscheidungen treffen (Öktem, 2022; Litzel & Luber, 2019).

Weiterführend können diese Systeme Voreingenommenheiten aufweisen, welche in diskriminierenden Entscheidungen resultieren können. Aus diesem Grund wird im Folgenden die Problematik der *algorithmischen Voreingenommenheit* vorgestellt und im weiteren Verlauf dieser Arbeit vertieft (Koval, 2020).

### **3.4 Bewusste- sowie unbewusste Voreingenommenheit**

Bevor genauer auf die Voreingenommenheit von Algorithmen eingegangen wird, wären zum Verständnis drei Begrifflichkeiten zu klären, die mit diesem Thema zusammenhängen und eine genaue Definition sowie Erläuterung verlangen.

Angefangen mit dem Begriff *Voreingenommenheit* oder im eng. *Bias*. Wenn ein Mensch sich entscheidet, eine Person oder eine Sache zu unterstützen oder sich dieser entgegenzusetzen und diese Entscheidung ungerechterweise von persönlicher Meinung beeinflusst wird, so spricht man von Voreingenommenheit bzw. einer voreingenommenen Entscheidung (*Bias*, o. D.; Gasser, 2019).

Im weiteren Verlauf dieser Arbeit wird Voreingenommenheit, Verzerrung sowie Vorurteil als Synonym betrachtet.

Hierbei kann zwischen *bewusster Voreingenommenheit* und *unbewusster Voreingenommenheit* unterschieden werden.

Bei der bewussten Voreingenommenheit ist der voreingenommenen Person vollkommen klar, dass sie eine vorurteilsbehaftete Meinung hat. Die Person würde diese jedoch trotzdem verfolgen, da sie diese Meinung wirklich vertritt. Beispielsweise könnte ein Arbeitgeber A1 explizit die Meinung vertreten, dass für eine Führungsposition nur Personen, die groß sowie männlich sind, in Betracht gezogen werden sollten (Bär, 2022; Bönisch, 2010; Burch, 2022).

Im Gegensatz hierzu wird eine Entscheidung, die auf unbewusster Voreingenommenheit basiert, nicht von der entscheidenden Person wahrgenommen. Bezieht man dies auf das Beispiel, so würde dies bedeuten, dass ein anderer Arbeitgeber A2 einen großen und männlichen Bewerber für die Führungsposition einstellt, jedoch selbst der Meinung ist, dass dieser Bewerber lediglich die beste Qualifikation von allen Bewerbern aufweist und er somit eine faire und neutrale Entscheidung fiel. Unterbewusst ist es jedoch so, dass A2 das Vorurteil, dass das männliche Geschlecht sich eher als Führungskraft eignet, verfolgt (Bär, 2022; Bönisch, 2010; Burch, 2022).

Weiterführend könnte A2 die Größe des Bewerbers mit Erfolg assoziiert haben, da größere Menschen öfter in Führungspositionen vertreten sind (Bär, 2022; Burch, 2022).

Im Ergebnis hat A2 ebenfalls anhand von Merkmalen, welche keinerlei Zusammenhang mit der tatsächlichen Qualifikation des jeweiligen Bewerbers für die gegebene Position haben, seine endgültige Entscheidung getroffen. Der Unterschied zwischen A1 und A2 besteht alleinig darin, dass A1 bewusst voreingenommen ist und somit explizit männliche Bewerber bevorzugt. Während A2 unbewusst voreingenommen ist und somit implizit diese Art von Bewerber bevorzugt. A2 ist hierbei ebenfalls nicht bewusst, dass er im Grunde eine voreingenommene Entscheidung fiel, sondern er ist der Meinung, dass seine Entscheidung vollkommen vorurteilsfrei war (Bär, 2022; Burch, 2022).

Voreingenommenheit resultiert aus historisch- sowie sozial angeeigneten Werten und Denkweisen einer Gesellschaft. Die unbewusste Voreingenommenheit ist aus menschenpsychologischen Gründen die häufiger vorkommende Variante. Außerdem weist diese für den weiteren Verlauf dieser Arbeit ebenfalls die meiste Relevanz auf, sodass auf diese im späteren Kapitel näher eingegangen wird (Bär, 2022; Burch, 2022).

### **3.5 Algorithmische Voreingenommenheit**

Da nun klar ist, was mit Voreingenommenheit sowie bewusster- und unbewusster Voreingenommenheit zu verstehen ist, kann nun explizit darauf eingegangen werden, was unter dem Begriff *algorithmische Voreingenommenheit* zu verstehen wäre (Koval, 2020).

Ein Verwendungsgrund von KI-Systemen bzw. AES ist, dass diese neutrale Entscheidungen fällen. Nun ist es jedoch so, dass sie oftmals aufgrund verschiedenster Ursachen dazu neigen, sich Voreingenommenheiten anzutrainieren. Diese antrainierten Vorurteile könnten dann in diskriminierenden Entscheidungen münden (Kilian, 2022; Koval, 2020).

Beispielsweise könnte man sich ein AES vorstellen, welches darüber entscheiden soll, wer von den Bewerbern sich für eine freie Position in einer Firma am besten eignet (De Montis, 2019).

Angenommen es haben sich 5 Frauen und 5 Männer mit ähnlicher Qualifikation beworben und das AES liefert exakt 5 Männer als „am besten geeignet“ zurück. Dann war die Entscheidungsfindung des Systems wahrscheinlich nicht komplett objektiv, sondern es ist vielmehr möglich, dass zusätzlich zu den qualifizierenden Eigenschaften das Geschlecht mit in die Prognosefindung berücksichtigt wurde. Dies würde erklären, wieso alle Männer und keine Frau von diesem System empfohlen wird, obwohl alle eine ähnliche Qualifikation aufweisen (De Montis, 2019).

Dieses Beispiel illustriert ein AES, welches möglicherweise *geschlechterspezifische Voreingenommenheit* aufweist und somit Frauen-diskriminierende sowie Männer-bevorzugende Entscheidungen trifft (De Montis, 2019; Machill, 2020).

### **3.6 Stereotypisierung**

Ein weiterer mit dem Beispiel der Voreingenommenheit bereits angeschnittener relevanter Begriff wäre die *Stereotypisierung*.

Diese kognitive Konstruktion entsteht aus der Gesellschaft und ihrer bestehenden Weltanschauung heraus. Stereotypisieren bedeutet im Wesentlichen einen Menschen auf Grundlage seines Geschlechtes, seines Alters, seiner Herkunft, seiner Religion, seiner Hautfarbe usw. in soziale

Gruppen einzuordnen (Kolleck & Orwat, 2020; Machill, 2020; Schulz, 2020; Six-Materna & Six, o. D.).

Diese Merkmale, welche zum Einordnen herangezogen werden, werden auch *geschützte Attribute* genannt. Die Gesellschaft kategorisiert Menschen bestimmter geschützter Attribute in soziale Gruppen ein, sodass die in der Gesellschaft lebenden Menschen diesen Stereotypisierungen ausgesetzt sind (Kolleck & Orwat, 2020; Machill, 2020; Schulz, 2020; Six-Materna & Six, o. D.).

Das menschliche Gehirn zieht sich einen Vorteil aus diesen Stereotypisierungen. Dies kann man sich so vorstellen, dass Menschen oftmals unterbewusst Entscheidungen auf Grundlage dieser Stereotypen treffen, da das Gehirn diese heranzieht, um ein schnelleres Urteil über die jeweilige Problemstellung oder Situation fällen zu können (Machill, 2020; Schulz, 2020; Six-Materna & Six, o. D.).

Ein Beispiel hierzu könnte für mehr Klarheit sorgen. Dafür wird ein bereits existierender Stereotyp herangezogen. Dieser wäre „Deutsche sind immer pünktlich [...]“ (Glas, 2018). Man stelle sich nun beispielhaft die Situation vor, dass ein Mensch ein Taxi bestellt und dabei zwischen zwei Fahrern wählen kann. Hierbei wäre der eine Fahrer ausländischer Herkunft und der andere Fahrer deutscher Herkunft. Sollte die Person nun den deutschen Fahrer wählen, so kann es sein, dass kognitiv unterbewusst der angesprochene Stereotyp in die Entscheidungsfindung einbezogen wurde (Machill, 2020; Schulz, 2020; Six-Materna & Six, o. D.).

Dies hat zwei Gründe. Der erste Grund wäre, dass das Gehirn dieser Person diesen Stereotyp zu der gegebenen Problemstellung mitberücksichtigt, da es relevant ist, wann der Fahrer die Person antrifft und mitnimmt. Der zweite Grund wäre, dass sich das Gehirn durch das Heranziehen dieser „Informationen“ Zeit und Aufwand für die Lösungssuche spart (Machill, 2020; Schulz, 2020; Six-Materna & Six, o. D.).

### **3.7 Diskriminierung**

Der nächste für den Kontext relevante Begriff wäre *Diskriminierung*.

Diese Form von Menschenrechtsverletzung betrifft alltäglich Millionen von Menschen. Wie bereits erläutert, kategorisiert bzw. stereotypisiert die Gesellschaft Menschen auf Grundlage ihrer geschützten Merkmale in soziale Gruppen (*Discrimination and Intolerance*, o. D.).

Nun spricht man von Diskriminierung, wenn Menschen auf Grundlage dieser sozialen Gruppen in einer gleichen Situation ungleich behandelt werden (*Discrimination and Intolerance*, o. D.; Fiegele, 2021; Machill, 2020).

Beispielsweise, wenn bei einer Jobausschreibung ein Bewerber X mit der gleichen Qualifikation wie ein Bewerber Y den Job nicht zugesichert bekommt, aber Y schon.

Wenn der Arbeitgeber den Bewerber X in eine weniger angesehene soziale Gruppe bzw. Minderheit kategorisiert hat und auf Grundlage dieser Kategorisierung ihm nicht den Job zugesichert hat, dann spricht man von Diskriminierung eines Individuums, in diesem Falle des Bewerbers X (*Discrimination and Intolerance*, o. D.; Fiegele, 2021; Machill, 2020).

Diskriminierung lässt sich also auf das Urteilen über Menschen auf Grundlage von geschützten Merkmalen zurückführen, die keinerlei Zusammenhang mit der Individualität und der tatsächlichen Qualifikation des Menschen zu tun haben (*Discrimination and Intolerance*, o. D.; Fiegele, 2021; Machill, 2020).

Diskriminierung beruht des Öfteren auf der Voreingenommenheit eines Menschen oder einer Gesellschaft. Des Weiteren ist Diskriminierung in vielen Sektoren unserer Gesellschaft präsent und kann schwere Folgen, wie mit dem soeben vorgestellten Beispiel illustriert, oder noch fatalere Folgen nach sich ziehen (*Discrimination and Intolerance*, o. D.).

Beispielsweise könnte es passieren, dass eine Person, die seit Jahrzehnten in einem Land lebt, eventuell aufgrund seiner geschützten Attribute, keine Staatsbürgerschaft gewährt bekommt (*Discrimination and Intolerance*, o. D.).

Diskriminierung kann in *direkte- und indirekte Diskriminierung* unterteilt werden (*Discrimination and Intolerance*, o. D.).



### **3.7.1 Direkte Diskriminierung**

Diese Form von Diskriminierung lässt sich daran erkenntlich machen, dass bewusst und gewollt auf Grundlage der wahrgenommenen sozialen Gruppenzugehörigkeit eines Menschen, also auf Basis seiner geschützten Attribute, über ihn geurteilt und gehandelt wird. Beim dem soeben vorgestellten Beispiel der Jobausschreibung wurde der Bewerber X direkt diskriminiert, da der Arbeitgeber ihn bewusst aufgrund von bestimmten geschützten Merkmalen, wie z. B. aufgrund seiner Hautfarbe und Herkunft, nicht eingestellt hat (*Discrimination and Intolerance*, o. D.; *Formen der Diskriminierung*, 2020).

### **3.7.2 Indirekte Diskriminierung**

Diese Form spiegelt sich unter anderem in vorerst harmlos wirkenden Regeln wie Mindestgrößen für Polizisten oder der Vorschrift, dass man keine Kopfbedeckung wie Hüte o. ä. auf der Arbeit tragen darf, wider (*Discrimination and Intolerance*, o. D.; *Formen der Diskriminierung*, 2020).

Doch diese Regeln betreffen manche Individuen stärker als andere, sodass die erste Vorschrift, das geschützte Attribut der Körpergröße diskriminiert, da Menschen unter 1.60m für den Beruf nicht zugelassen werden (*Einstellungsverfahren*, o. D.). Berücksichtigt man nun die Tatsache, dass Frauen statistisch gesehen kleinwüchsiger sind als Männer, so hat man zum Ergebnis eine weitere Diskriminierung vorliegen, nämlich gegenüber dem geschützten Merkmal des Geschlechtes, genauer, dem weiblichen Geschlecht (*Discrimination and Intolerance*, o. D.; *Formen der Diskriminierung*, 2020; *Verteilung der Körpergrößen nach Geschlecht im Jahr 2006*, 2006).

Die zweite Regel würde das geschützte Attribute der religiösen Angehörigkeit eines Menschen verletzen. In diesem Falle werden insbesondere muslimische Frauen, die aufgrund ihres Glaubens eine Kopfbedeckung tragen wollen, diskriminiert, da sie diese an dem jeweiligen Arbeitsort nicht tragen dürfen (*Discrimination and Intolerance*, o. D.; *Formen der Diskriminierung*, 2020).

Abschließend ist wichtig zu erwähnen, dass die indirekte Form von Diskriminierung im alltäglichen Leben präsenter und darüber hinaus schwieriger zu entdecken sowie nachzuweisen ist als die direkte Form (*Discrimination and Intolerance*, o. D.).

### **3.8 Fairness**

An dieser Stelle wäre der komplementäre Begriff *Fairness* zu erläutern.

Dieser versteht sich als Gegensatz zur Diskriminierung, sodass zur Erfüllung der Fairness bei Entscheidungen, wie z. B. ob der Bewerber X oder Bewerber Y eingestellt werden sollte, lediglich auf die Individualität und Qualifikation der Person geachtet werden sollte. Demzufolge muss zur Erreichung von Fairness, die Voreingenommenheit sowie die damit im Zusammenhang stehende Diskriminierung abwesend sein. Dies würde im Kern bedeuten, dass über Menschen nicht auf Grundlage ihrer geschützten Attribute geurteilt wird, sondern vielmehr auf Grundlage von objektiven Merkmalen (Fiegele, 2021).

Darüber hinaus lassen sich hier ebenfalls zwei Arten unterscheiden, nämlich *Objektive-Fairness* und *Subjektive-Fairness*. Objektive Fairness lässt sich messen und testen, während subjektive Fairness von Individuum zu Individuum variiert. Dies würde bedeuten, dass wenn Bewerber Y die Stelle zugewiesen bekommt und Bewerber X nicht, obwohl beide die exakt gleichen Qualifikationen aufweisen, dann könnte es sein, dass jemand dies als fair, während jemand anderes dies als unfair ansieht (Fiegele, 2021).

## 4 Vorteile von KI-Systemen bzw. AES

Bevor im nächsten Kapitel mit der allumfassenden Ausführung der Thesis begonnen wird, sollte vorerst verdeutlicht werden, welche Vorteile die Verwendung von KI-Systemen bzw. AES mit sich bringt.

Dieses Kapitel soll verdeutlichen, wieso die Integrierung von AES so reizvoll ist für viele Unternehmen und dass diese Systeme trotz deren eher kritischen Betrachtung in den folgenden Kapiteln viele Vorteile und somit Verwendungsgründe aufweisen.

Der Autor dieser Arbeit fokussiert sich auf die seiner Meinung nach fünf wichtigsten Vorteile.

Angefangen mit der *Effizienz*: KI-Systeme können Unmengen an Daten effizient verarbeiten und somit rapider Entscheidungen fällen als Menschen. Dementsprechend können AES ebenfalls deutlich schneller beispielsweise Daten zu einer Person analysieren und darüber entscheiden, ob diese einen Kredit erhalten sollte oder nicht. Die Fähigkeit effizienter zu arbeiten, ist eines der Kernvorteile von KI und somit für viele Unternehmen ein Verwendungsgrund (Künstliche Intelligenz Vor- und Nachteile, 2023).

Weiter mit der *Genauigkeit*: Aufgrund der Tatsache, dass AES ihr Verhalten auf einer aufbereiteten großen Datenmenge erlernen und, wie soeben erörtert, sie große Datenmengen effizient verarbeiten können, haben sie die Fähigkeit, in ihrem erlernten Bereich, sehr Präzise Ergebnisse hervorzubringen (Künstliche Intelligenz Vor- und Nachteile, 2023).

Außerdem sind KI-Systeme in der Lage, *diverse komplexe Aufgaben* zu bewältigen. Aufgrund der erhöhten Effizienz sowie Genauigkeit liegt es nahe, bei solchen Aufgaben, für wessen Bewältigung Menschen eventuell mehr Zeit benötigen würden, somit ebenfalls auf die Verwendung von KI zurückzugreifen (Clickblogger, 2022).

Weiterführend sind AES rund um die Uhr einsatzfähig und benötigen im Optimalfall keinen Menschen, welcher diese Systeme anstößt. Also liegen mit KI-Systemen automatisierte, jederzeit verfügbare sowie produktiv arbeitende Systeme vor (Kilian, 2022).

Abschließend werden getroffene Entscheidungen sowie Vorhersagen von KI-Systemen bzw. AES oftmals als *neutrale Wahrheit* anerkannt. Dies hat den Grund, dass angenommen wird, dass KI-Systeme auf Grundlage von objektiven Daten sowie eines objektiven Verhaltens, leistungsfähig, korrekt sowie neutral arbeiten. Also wird davon ausgegangen, dass KI-Systemen eine effiziente sowie unvoreingenommene Entscheidungsfindung zugrunde liegt und somit die von diesen Systemen erzeugten Ergebnisse, ohne Einbeziehung von Vorurteilen, hervorgebracht werden (De Montis, 2019; Kilian, 2022).

Die kontroverse Annahme, dass KI-Systeme bzw. AES auf objektiver Grundlage Ergebnisse erzeugen, wird im weiteren Verlauf dieser Arbeit kritisch thematisiert, da diese Annahme nicht realitätsgetreu ist. Im Folgenden wird dementsprechend insbesondere darauf eingegangen, wie Voreingenommenheiten in diese KI-Systeme gelangen, welche möglicherweise verheerenden Auswirkungen dies mit sich bringt und wie man die in diesen Systemen vorhandenen Vorurteile mindern kann.

Wichtig hervorzuheben wäre, dass diese Arbeit versucht, einen allumfassenden Blick auf algorithmische Voreingenommenheit zu ermöglichen. Aus diesem Grund wird diese Thematik in seinen Ursprüngen angegangen, wodurch im nächsten Kapitel vorerst auf die Ursachen menschlicher kognitiver Verzerrung eingegangen wird.

## 5 Ursachen kognitiver Verzerrung

Im Folgenden wird vorgestellt, wie Menschen zu einem vorurteilsbehafteten Meinungsbild gelangen. Denn mit Hintergrund dieses Wissens kann im späteren Verlauf dieser Arbeit hervorgehoben werden, wie dieses voreingenommene Meinungsbild sich auf KI-Systeme bzw. AES übertragen kann.

Zuallererst ist wichtig zu wissen, dass die Mehrheit der Denkprozesse unterbewusst ablaufen, sodass Entscheidungen, welche man als bewusste wahrnimmt, häufig entweder zum Teil oder komplett auf Grundlage von unterbewussten Denkvorgängen entstehen (Bär, 2022; Rosso, 2018).

Dies hat den Grund, dass effizientes Arbeiten für unser Gehirn besonders wichtig ist, da es trotz seiner geringen Masse trotzdem 20% unserer Körperenergie aufbraucht, wobei der Großteil dieser Energie für das *bewusste Denken* gebraucht wird (Bär, 2022; Rosso, 2018).

Demzufolge benötigt und besitzt unser Gehirn Methodiken wie das *unterbewusste Denken*, um diesen enormen Energieverbrauch zu mindern und die ineffiziente Weise des bewussten Denkens so weit zu reduzieren, wie nur möglich (Bär, 2022; Rosso, 2018).

Hierfür zieht unser Gehirn Vorurteile mit in das unterbewusste Denken ein. Mithilfe dieser kann unser Gehirn effizienter und schneller Ergebnisse hervorbringen, da seine Denkvorgänge unter Berücksichtigung von *kognitiver Verzerrung* unter anderem gedankliche Abkürzungen suchen, um zügiger und somit energiesparender Lösungen hervorzubringen (Bär, 2022; Rosso, 2018).

Es existiert eine Vielzahl an Vorurteilen, die ein menschliches Gehirn heranziehen könnte, sodass im Folgenden die Ursachen kognitiver Verzerrung in vier Kategorien niedergelegt und zu diesen dann einige für den weiteren Verlauf relevante entstehende Verzerrungen zugeordnet werden (Bär, 2022; Gasser, 2019; Rosso, 2018).

## **5.1 Übermaß an Informationen**

Dadurch, dass uns tagtäglich eine für unser Gehirn nicht handhabbare Menge an Informationen zur Verfügung gestellt wird, filtert unser Gehirn den Großteil dieser Informationen aus und persistiert nur diejenigen Informationen, von welchen es der Meinung ist, dass diese später einmal relevant sein könnten (Benson, 2016; Gasser, 2019).

Diese möglicherweise relevanten Informationen, welche eher gespeichert werden, wären folgende:

### **5.1.1 Informationen, die mit bestehenden Informationen zusammenhängen**

Die hieraus resultierende Verzerrung wäre die *Verfügbarkeitsverzerrung*. Bei dieser Verzerrung würde das Gehirn Erfahrungen aus einem vergangenen Ereignis exakt auf das aktuell vorhandene sowie eventuell veränderte Ereignis übertragen, wodurch ein Denkfehler vorliegt (Bär, 2022; Benson, 2016; Stappert, 2022).

Ein Beispiel hierfür könnte ein Geschehen sein, bei welchem ein Ausländer eine Bank ausraubt. Menschen, die dieses Ereignis miterlebt haben, würden nun möglicherweise beim Treffen eines anderen Ausländers dieses Geschehen mit dem Treffen verknüpfen und durch die Verfügbarkeitsverzerrung davon ausgehen, dass diese Person ebenfalls kriminell ist (Bär, 2022; Benson, 2016; Stappert, 2022).

### **5.1.2 Informationen, welche bestehende Informationen aktualisieren**

Die hier entstehende Verzerrung wäre die Verzerrung namens *Ankerheuristik*. Bei dieser nimmt sich das Gehirn zu einer gegebenen Problemstellung, zu welcher er die Antwort nicht kennt, eine Information als Ausgangsposition bzw. Anker her und versucht, anhand dessen die Lösung für die Problemstellung zu schätzen (Balthasar, 2021; Benson, 2016; Stappert, 2022).

Dieser Anker kann entweder mit der Problemstellung zusammenhängen oder komplett willkürlich und irrelevant sein. Beispielsweise könnte man gefragt werden, wer als Täter einer

Schießerei in Frage kommen könnte und während einem diese Frage gestellt wird, hört man Musik, in welcher eine afroamerikanische Person über Waffen und Gefahren rappt. Dann könnte das Gehirn die für die Problemstellung irrelevanten und willkürlichen Informationen als Anker heranziehen und demzufolge als mögliche Täter afroamerikanische oder dunkelhäutige Personen verdächtigen (Balthasar, 2021; Benson, 2016; Stappert, 2022).

### **5.1.3 Informationen, welche bestehende Meinung bestätigen bzw. verstärken**

Hieraus resultiert der *Bestätigungsfehler* oder im eng. *Confirmation Bias*. Bei dieser Verzerrung spielt das bereits vorhandene Meinungsbild eines Menschen eine große Rolle. Informationen, welche dem bestehenden Meinungsbild widersprechen, tendieren eher dazu, ignoriert bzw. als falsch eingeordnet zu werden, im Gegensatz zu Informationen, die das vertretene Meinungsbild validieren und somit als richtig eingeordnet werden (s. Abb. 5) (Bär, 2022; Benson, 2016; *Confirmation Bias: Warum der Bestätigungsfehler so gefährlich ist*, o. D.).

Demzufolge fallen einem bei dieser Verzerrung auch eher Informationen auf, die das vertretene Meinungsbild bestätigen. Ein Beispiel für diese Verzerrung könnte man anhand des bereits herangezogenen Stereotypen „Deutsche sind immer pünktlich“ näherbringen. Sollte eine Person diesen Stereotypen verinnerlicht haben und würde diese nun zwei Artikel darüber lesen, einen, dass statistisch gesehen Deutsche unpünktlich sind, und einen Artikel darüber, dass Deutsche, nach der Erfahrung des jeweiligen Autors, pünktlich sind, dann würde der Leser im Sinne des Bestätigungsfehlers den Stereotypen auflösenden Artikel außer Acht lassen und den Stereotypen befürwortenden Artikel berücksichtigen und im Gehirn speichern. Erkenntlich wird, dass bei vorhandenem Bestätigungsfehler einerseits sogar Fakten außen vor gelassen werden, um die bestehende Meinung beizubehalten, und andererseits, dass mit dieser Verzerrung mögliche bestehende diskriminierende Meinungsbilder verstärkt werden könnten (Bär, 2022; Benson, 2016; *Confirmation Bias: Warum der Bestätigungsfehler so gefährlich ist*, o. D.).

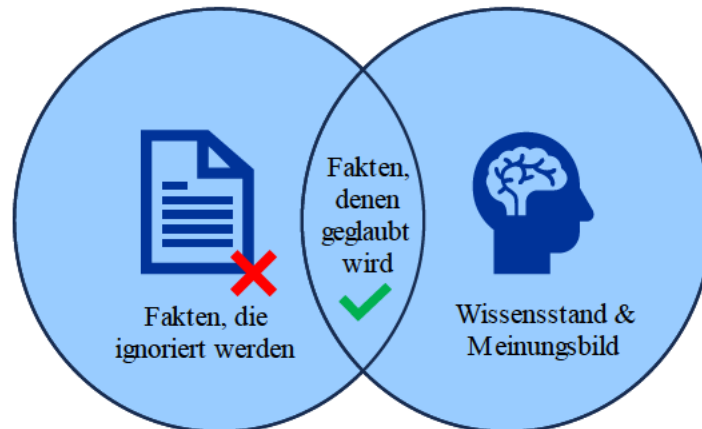


Abbildung 5: Confirmation Bias (eigene Darstellung)  
(Confirmation Bias: Warum der Bestätigungsfehler so gefährlich ist, o. D.)

#### 5.1.4 Menschen erkennen eigene Denkfehler nicht

Hiermit ist die *Verzerrungsblindheit* oder im eng. der *Blind-Spot-Bias* gemeint. Dieser Gedankenfehler besagt, dass Menschen ihr eigenes Gedankenbild als sinnvoll und selbstbestimmt wahrnehmen, während dies, wie mit diesem Kapitel verdeutlicht, nicht der Fall sein kann, da das Gehirn primär für unterbewusste Denkprozesse und zur Steigerung der Effizienz Verzerrungen heranzieht (Benson, 2016; Stappert, 2022).

## 5.2 Nicht genügend Bedeutung

Anhand der reduzierten Menge an Informationen, die ein Mensch persistieren kann, versucht das Gehirn nun, diese miteinander zu verknüpfen und ein gesamtheitlich „sinnvolles“ Weltbild zu formen. Hierbei werden mögliche Informationslücken mit Erfahrungswerten gefüllt, welche unser Gehirn für die jeweilige Lücke als sinnvoll empfindet (Benson, 2016; Gasser, 2019).

Die für unser Gehirn zur Füllung der Lücken relevanten Informationen wären folgende:



### **5.2.1 Informationen, welche als Lückenfüller fungieren**

Informationen, wie beispielsweise Stereotype, werden herangezogen, wenn Informationslücken bestehen. Dabei werden diese Lückenfüller mit den bereits bestehenden Informationen gleichgesetzt, sodass unser Gehirn nichtmehr unterscheiden kann, was die tatsächlichen evtl. richtigen Informationen waren und welche hinzugedacht wurden (Benson, 2016).

### **5.2.2 Informationen, die zu einer Person fehlen, würden hinzugedacht werden**

Hierbei wird jedoch unterschieden zwischen Menschen, die wir kennen oder mögen, und Menschen, die wir nicht kennen oder mögen. Die positiven Attribute der uns bekannten oder gemochten Personen werden von unserem Gehirn auf die weiteren unbekannt Attribute bzw. Informationslücken übertragen. Weiterführend lassen wir hier ebenfalls miteinspielen, welche Werte und Denkweisen wir persönlich vertreten und würden diese dann ebenfalls für die Informationslückenfüllung nutzen, da wir davon ausgehen, dass diese Person ein ähnliches Weltbild verfolgt wie wir. Gegensätzliches gilt dann für diejenigen Menschen, die wir nicht kennen oder mögen (Benson, 2016; Stappert, 2022).

Diese Verzerrung wird auch *Halo-Effekt* genannt. Ein Beispiel für diese Verzerrung könnte ein Mensch sein, welcher grundsätzlich ausländerfeindlich ist und nun Aussagen darüber treffen muss, ob Person x (Herkunft: Türkei) kriminell oder Person y (Herkunft: Deutschland) kriminell ist. Dann würde dieser Mensch bzw. sein Gehirn aufgrund des Halo-Effektes die negativen Assoziationen, die er mit Person x verbindet, und die positiven, die er mit Person y verbindet, nutzen und auf diese zu schätzende Informationslücke übertragen. Zum Ergebnis würde der Mensch zu dem Entschluss kommen, dass Person x kriminell und Person y nicht kriminell ist (Benson, 2016; Stappert, 2022).

Abgesehen vom Halo-Effekt existiert eine weitere hierzu relevante Verzerrung namens *In-Group bias* oder *gruppeninterne Verzerrung*. Bei dieser Verzerrung würden Mitglieder einer sozialen Gruppe Angehörige dieser bevorzugen, während sie Mitglieder fremder Gruppen abwerten würden. Dies tun sie mit dem Ziel, die positive Identität der Gruppe aufrecht zu erhalten. Dies kann dann eine große Rolle spielen, wenn beispielsweise eine Frau in einem

Unternehmen angestellt sein möchte, welches zu 90% aus Männern besteht, und der Arbeitgeber dieses Unternehmens aufgrund der gruppeninternen Verzerrung diese Frau nicht einstellt, da sie kein Mann ist (Fiegele, 2021).

### **5.3 Die Notwendigkeit, schnell zu handeln**

Die zu treffende Entscheidung eines Menschen für eine gegebene Situation ist durch Zeit sowie Informationen eingeschränkt. Demzufolge ist jede neu dazukommende Information rapide zu bewerten und in die Entscheidungsfindung einzubeziehen (Benson, 2016; Gasser, 2019).

Um schnelle Entscheidungen treffen zu können, enthält unser Gehirn folgende Eigenschaften:

#### **5.3.1 Die Eigenschaft, zu handeln**

Die Eigenschaft zu handeln. Diese wird benötigt, um überhaupt eine Entscheidung treffen zu können. Hierfür muss man selbstbewusst und zuversichtlich in seinem Urteil sein. Dieses Selbstbewusstsein ist des Öfteren Ursache von Selbstüberschätzung, daher findet hier unter anderem die Verzerrung namens *Overconfidence effect* oder *Effekt der Selbstüberschätzung* Anwendung (Benson, 2016; MasterClass, 2022).

Diese Verzerrung würde dafür sorgen, dass Menschen ihre Fähigkeiten sowie ihr Wissen überschätzen und im Grunde empfinden, dass sie „alles können“ und „alles wissen“. Dementsprechend haben Menschen, die dieser Verzerrung zum Opfer gefallen sind, eine eingeschränkte Entscheidungsfindung, da unter anderem ihre objektive Sicht des Sachverhaltes kaum vorhanden ist. Diese Verzerrung verstärkt und erweitert diskriminierende Vorurteile und Stereotypen eines Menschen. Darüber hinaus könnte die selbstbewusste Art und Weise der Überbringung dieser Ansichten einen anderen Menschen ebenfalls beeinflussen und ihn somit von diesen eventuell diskriminierenden Werten überzeugen (Benson, 2016; MasterClass, 2022).

### **5.3.2 Die Eigenschaft, fokussiert zu bleiben**

Die Eigenschaft fokussiert zu bleiben. Diese wird benötigt, um kontinuierlich an der Entscheidungsfindung arbeiten zu können. Hierbei priorisiert unser Gehirn Aufgaben und Problemstellungen, die aktuell anliegen, höher als diejenigen, die in der Zukunft zu erledigen sind (Benson, 2016).

### **5.3.3 Die Eigenschaft, Fehler vermeiden zu wollen**

Als Mensch wollen wir Fehlentscheidungen vermeiden, damit unsere Reputation innerhalb einer sozialen Gemeinschaft nicht fällt. Hier findet unter anderem die *Status-quo-Verzerrung* Anwendung. Menschen würden bei dieser Verzerrung unter anderem aufgrund von Verlustängsten, jegliche Entscheidungen dahingehend treffen, ihren aktuellen Zustand aufrecht zu erhalten (Benson, 2016; Engelkamp, o. D.).

Demzufolge wehren sich diese Menschen gegenüber Veränderungen und wollen diese möglichst vermeiden. Hier kann man als Beispiel eine Person heranziehen, die ein diskriminierendes Weltbild gegenüber Ausländern aufrechterhalten möchte. Diese Person würde ihr bestmögliches versuchen, ihre Entscheidungen dahingehend so zu treffen, dass ihre Vorurteile bestehen bleiben. Demzufolge könnten sie beispielsweise zum Ergebnis eine rechte Partei wählen (Benson, 2016; Engelkamp, o. D.).

## **5.4 Verallgemeinerungen werden eher gespeichert als Details**

Das Gehirn nimmt täglich eine riesige Menge an Informationen auf und muss dann entscheiden, welche Informationen es verwerfen sollte oder welche persistiert werden sollten (Benson, 2016; Gasser, 2019).

Diese Problemstellung wurde bereits erörtert und es wurde ebenfalls illustriert, wie unser Gehirn entscheidet, welche Informationen er eher speichert und welche er eher verwirft.

Nichtsdestotrotz lässt sich sagen, dass wir Verallgemeinerungen eher und vereinfachter speichern können als Details und Einzelheiten (Benson, 2016; Gasser, 2019).

Demzufolge existieren folgende weitere Eigenschaften des Gehirns:

#### **5.4.1 Die Eigenschaft der Verallgemeinerung**

Also spezifische Informationen zu verwerfen und das große ganze Allgemeingültige zu speichern. Der hiermit verbundene *Negativity bias* oder die *Negative Voreingenommenheit*, welche beschreibt, dass negative Erlebnisse oder Erkenntnisse stärker verinnerlicht werden als positive, mündet darin, dass bei einer Person Vorurteile oder Voreingenommenheiten gegenüber einer anderen Person oder Gruppe entstehen können, obwohl diese Person mehr positive als negative Erfahrungen mit dem jeweiligen Gegenüber erfuhr. Darüber hinaus könnte dies in einem allgemeingültigen und generellen Vorurteil gegenüber dieser anderen Person oder Gruppe münden (Benson, 2016; Loeffner, o. D.).

Beispielsweise könnte eine Person viele positive Erfahrungen mit Ausländern gehabt haben, jedoch trifft sie ab und an auch mal Ausländer, mit welchen sie negative Erfahrungen erfährt. Nun könnte das Gehirn den Negativity Bias sowie weitere Stereotypen und Vorurteile in der Verallgemeinerung von Ausländern mitberücksichtigen, sodass im Ergebnis ein voreingenommenes und ausländerfeindliches Meinungsbild vorliegt (Benson, 2016; Loeffner, o. D.).

#### **5.4.2 Die Eigenschaft, Erinnerungen zu verstärken**

Hiermit ist gemeint, dass unser Gehirn das Sichern von einigen Erinnerungen verstärkt, während dabei Einzelheiten dieser vertauscht oder neue Details hinzugefügt werden können. Ein hiermit im Zusammenhang stehender Begriff wäre die *Gedächtnistäuschung*. Bei dieser Täuschung würden wir falsche Erinnerungen aufrufen, also Erinnerungen, von welchen wir überzeugt sind, dass diese so passiert und richtig sind, während sie in Wirklichkeit so nie passiert und dementsprechend falsch sind (Benson, 2016; *Gedächtnistäuschung*, o. D.).

Mit diesem Kapitel wurde nähergebracht, dass aus Effizienzgründen die meisten unserer Denkprozesse unterbewusst ablaufen. Die mit diesen Prozessen im Zusammenhang stehenden kognitiven Verzerrungen wurden kategorisiert und näher erläutert.

Hierbei wurden diejenigen Verzerrungen, welche ein diskriminierendes sowie voreingenommenes Meinungsbild zum Ergebnis haben können, hervorgehoben. Insbesondere die diskriminierenden Beispiele, wie sie in der realen Welt aufkommen könnten, haben die Relevanz dieser Verzerrungen für den weiteren Verlauf dieser Arbeit hervorgehoben.

Wichtig ist hierbei zu berücksichtigen, dass unser Gehirn diese Verzerrungen nicht isoliert, sondern im Zusammenspiel ausnutzt. Insgesamt wurde somit verdeutlicht, dass auf Basis dieser unterbewussten verzerrten Denkprozesse ein vorurteilsbehaftetes sowie stereotypisierendes und im Ergebnis diskriminierendes und voreingenommenes Weltbild bei Individuen entstehen kann (Bär, 2022).

Die Art der unterbewussten kognitiven Verzerrungen sowie die damit im Zusammenhang stehenden Beispiele haben ebenfalls illustriert, wie die jeweilige Verzerrung aus einem voreingenommenen Meinungsbild resultieren könnte, bei welchem eine spezifische soziale Gruppe benachteiligt wird, ohne dass dies das bewusste Ziel der voreingenommenen Person war (Bär, 2022).

Dies liegt insbesondere an der Art und Weise, wie unser Gehirn arbeitet. Folglich können Menschen aufgrund ihrer ausgeprägten unterbewussten Denkprozesse und der in diesen Prozessen miteinspielenden kognitiven Verzerrungen als irrationales Lebewesen eingestuft werden, da sie Entscheidungen und Urteile eher auf Grundlage von Erfahrungswerten sowie Meinungsbildern fällen, anstatt primär objektive Daten zu berücksichtigen (Braswell, 2022).

Demzufolge sind ebenfalls insbesondere Menschen, die der Meinung sind, nicht voreingenommen zu sein, diejenigen, die unterbewusst ein voreingenommenes sowie evtl. diskriminierendes Weltbild verfolgen (Bär, 2022; Braswell, 2022).

## 6 Voreingenommenheiten bezüglich der geschützten Merkmale eines Menschen

Die bereits angesprochenen geschützten Merkmale eines Menschen werden durch diese kognitiven Verzerrungen angegriffen und verletzt. Voreingenommenheiten und die daraus resultierenden Verletzungen der geschützten Attribute sind das Hauptproblem von diskriminierenden AES, da sie in den Sektoren, in welchen AES meist Anwendung finden bereits vorherrschen. Diese Vorurteile lassen sich ebenfalls in Typen kategorisieren. Dementsprechend wären deren Einführung und nähere Erläuterung der nächste durchzuführende Schritt (Machill, 2020).

### 6.1 Rassistische Voreingenommenheit

Angefangen mit der *rassistischen Voreingenommenheit* oder im eng. *Racial Bias*. Bei dieser Voreingenommenheit würden Menschen unterbewusst vorurteilsbehaftete Assoziationen und Stereotypen in ihrer Entscheidung berücksichtigen und somit im Endresultat unwissentlich eine diskriminierende Entscheidung fällen (Braswell, 2022; Maryfield, 2018).

Dies ist nicht mit Rassismus gleichzusetzen, da eine rassistische Person bewusst der Meinung ist, dass eine Personengruppe über einer anderen steht. Demzufolge würde eine rassistische Person auch bewusst und wissentlich diskriminierende Entscheidungen treffen. Racial Bias beruht jedoch auf den bereits erörterten Verzerrungen, sodass eventuell verzerrte Erfahrungswerte sowie verzerrte Informationen im Gehirn gesichert wurden und nun zur Entscheidungsfindung berücksichtigt werden, sodass unwissentlich auf einer vorurteilsbehafteten rassistischen Grundlage die Lösung zu einer jeweiligen Problemstellung ermittelt wird (Braswell, 2022; Maryfield, 2018).

Trotz zahlreicher schockierender historischer Ereignisse existiert an vielen Orten der Welt täglicher Rassismus, ob der Hass gegenüber bestimmten Ethnien oder Personen bestimmter Religion, deren Diskriminierung ist gesetzlich verboten (Kolleck & Orwat, 2020).

Nichtsdestotrotz findet sich diese am Arbeitsplatz, im Gesundheitswesen, in sozialen Netzwerken usw. wieder und erreicht somit zahlreiche Menschen. Weiterführend würden diese Einflüsse sich dann unterbewusst in die Gedächtnisse der Menschen einschleusen, sodass ihr Gehirn bei Bedarf auf diese Diskriminierungen zugreifen kann (Machill, 2020).

## **6.2 Geschlechterspezifische Voreingenommenheit**

Der nächste vorurteilsbehaftete Typ wäre die *geschlechterspezifische Voreingenommenheit* oder im eng. der *Gender Bias*. Zur heutigen Zeit existieren viele verschiedene Auffassungen des Geschlechtes, jedoch beschränkt sich diese Arbeit, wenn nicht anders angegeben, die Geschlechter Mann und Frau (Machill, 2020).

Diese Voreingenommenheit fokussiert sich eher auf die Ungleichbehandlung des weiblichen und der Bevorzugung des männlichen Geschlechtes. Auch wenn die Geschlechter immer mehr gleichbehandelt werden, bestehen heutzutage immer noch benachteiligende Konzepte, Regeln sowie Vorurteile gegenüber der Frau (Machill, 2020).

Beispielsweise ermöglichen nur sechs Länder arbeitsrechtliche genderneutrale Gleichbehandlung. In den meisten Ländern wird die Frau in diesem Bereich benachteiligt behandelt. Dies spiegelt sich insbesondere in der Besetzung von Führungspositionen wider. Wie in einem vorangegangenen Beispiel bereits illustriert, neigen Arbeitgeber unterbewusst eher dazu, männliche Kandidaten für eine Führungsposition einzustellen als weibliche Kandidatinnen. Dies liegt daran, dass im gesellschaftlichen Bild, Männer eher als dominant und autoritär angesehen werden und diese Eigenschaften für eine Führungsposition als relevant eingestuft werden (Machill, 2020).

Außerdem werden Frauen im Bereich des Finanzsektors ebenfalls benachteiligt. Eine Untersuchung zeigte auf, dass Frauen in einigen Ländern, eine geringe Chance darauf haben, ein

Bankkonto zu besitzen und somit weniger wahrscheinlich Geld sparen, leihen oder anlegen werden (Machill, 2020).

Dies sind jedoch essenzielle Bestandteile, um finanziellen Reichtum sowie finanzielle Freiheit zu erlangen und weiterführend Eigentümerschaft zu erhalten (Machill, 2020).

Darüber hinaus können Frauen weitere Benachteiligung erfahren. Beispielsweise indem ihnen eine geringere Kreditwürdigkeit zugeschrieben wird oder sie grundsätzlich eine geringe Wahrscheinlichkeit zugeschrieben bekommen, ein Hypothekendarlehen zu erhalten (Machill, 2020).

Demzufolge werden Männer in vielen Bereichen bevorzugt und Frauen benachteiligt, jedoch gibt es auch Bereiche, in welchen genau das Gegenteil der Fall ist. Beispielsweise im strafrechtlichen Bereich. Hier erhalten bei gleicher Straftat Männer ein härteres Strafurteil als Frauen (Machill, 2020).

### **6.3 Altersvoreingenommenheit**

Eine weitere mögliche Form von Verzerrung liegt mit der *Altersvoreingenommenheit* vor. Dieses Attribut ist durch eine Richtlinie der Europäischen Union geschützt, welche besagt, dass innerhalb eines Beschäftigungsrahmens eine Gleichbehandlung aller Altersgruppen durchzusetzen ist (Machill, 2020).

Hiermit sollen insbesondere ältere Mitarbeiter geschützt werden, da die ältere Altersgruppe aufgrund von Stereotypen als langsamer, weniger anpassungsfähig, weniger trainingsbegeistert, weniger kreativ und anfälliger für Krankheiten angesehen wird. Abgesehen von der Tatsache, dass dieser Stereotyp (wie jeder andere auch) nicht zutreffen muss, wäre zudem noch wichtig zu erwähnen, dass ältere Mitarbeiter oftmals das meiste Fach- sowie Domänenwissen besitzen und somit eventuell effektiver arbeiten können als jüngere Mitarbeiter (Machill, 2020).

Weiterführend sind es oftmals die älteren Mitarbeiter, welche die jüngeren neu anfangenden Mitarbeiter einführen. Demzufolge sind diese von hoher Relevanz für das Unternehmen (Machill, 2020).



Nichtsdestotrotz zeigten verschiedenste Studien, dass insbesondere jüngere Mitarbeiter an altersdiskriminierenden Meinungsbildern gegenüber älteren Mitarbeitern festhalten, sodass sie es bevorzugen, mit Mitarbeitern unter 30 Jahren zusammenzuarbeiten. Auf der anderen Seite sind ältere Mitarbeiter in diesem Bereich weniger diskriminierend als ihre jüngeren Kollegen, sie arbeiten gleich gerne mit jüngeren oder älteren Kollegen zusammen (Machill, 2020).

Dies könnte dadurch begründet sein, dass ältere Mitarbeiter bereits viele Altersstufen im Leben durchlaufen konnten und somit Sympathie für die jüngeren Mitarbeiter aufbauen können. Grundsätzlich lässt sich also Altersdiskriminierung am Arbeitsplatz feststellen. Doch nicht nur dort herrscht eine Altersvoreingenommenheit (Machill, 2020).

Weitere Aufkommen sind besonders im Finanzsektor und im strafrechtlichen Bereich wiederzufinden. Im ersteren variiert die Akzeptanz eines Hypothekendarlehn-Antrags nach Altersstufe. Die Akzeptanzrate nimmt mit steigendem Alter zu, da wahrscheinlich Menschen mit zunehmendem Alter eine immer weiter steigende Wahrscheinlichkeit besitzen, einerseits überhaupt Vollzeit zu arbeiten, und andererseits einen möglichen Anstieg in der Karriereleiter aufweisen zu können, und somit eine höhere Chance besteht, dass sie liquide-technisch in der Lage sind, die Hypothek-Kosten zu tragen (Machill, 2020).

Diese Rate nimmt jedoch im erhöhten Alter wieder ab, dies liegt wohl daran, dass die Gefahr besteht, dass diese Menschen baldig in Rente gehen oder dass Menschen im höheren Alter anfälliger für lebensbedrohliche Krankheiten sind (Machill, 2020).

Im strafrechtlichen Bereich zeigt sich, dass jüngere Straftäter (bis zu 20 Jahren) und ältere Straftäter (ab 50 Jahren) mildere Urteile für dieselben Straftaten erfahren als Menschen innerhalb dieser beiden Grenzen. Die Schwere des Urteils nimmt für 20-30-Jährige stetig zu und weist seinen Höhepunkt für Straftäter im Alter von 30 Jahren auf. Nach dem 30. Lebensjahr nimmt die Schwere des Urteils kontinuierlich ab (Machill, 2020).

Hiermit wurde aufgezeigt, dass Altersdiskriminierung äußerst präsent ist in unserem alltäglichen Leben, da es in vielen Bereichen auftritt, mit welchen man zum Teil täglich in Berührung kommt (Machill, 2020).

## **6.4 Die Voreingenommenheit bezüglich sexueller Orientierung**

Ein Typ Vorurteil, welcher immer relevanter wird, wäre die *Voreingenommenheit in Bezug auf die sexuelle Ausrichtung* oder im eng. der *Sexual Orientation Bias*. Bei dieser Voreingenommenheit würden Menschen die sexuelle Ausrichtung einer Person mit in ihren Entscheidungen oder Handlungen berücksichtigen (Machill, 2020).

Hierfür hat die europäische Union ebenfalls Rechtsvorschriften erstellt, die Personen vor Diskriminierung auf Grundlage ihrer sexuellen Orientierung schützt (Machill, 2020).

Somit gehört die sexuelle Ausrichtung eines Menschen, wie alle in diesem Kapitel angesprochenen und anzusprechenden Attribute, ebenfalls zu einem der geschützten Merkmale (Kolleck & Orwat, 2020; Machill, 2020).

Diese Gesetzgebung bezieht sich jedoch nur auf den Bereich der Beschäftigung im Sinne eines Arbeitsverhältnisses, sodass Mitarbeiter dort vor dieser Art von Diskriminierung geschützt sein müssen (Machill, 2020).

Nichtsdestotrotz sorgt der alleinige Fokus dieser Gesetzgebung auf diesem bestimmten Bereich dafür, dass viele weitere Bereiche, in welchen diese Form von Diskriminierung ebenfalls aufkommen kann, gesetzlich unberührt bleiben. Viele dieser weiteren relevanten Gesetze sind bereits in *der Richtlinie zur Gleichbehandlung der Rassen* von der Europäischen Union niedergeschrieben, jedoch verfolgt nur ein Bruchteil der Mitgliedsstaaten diese Handlungsorientierung, sodass Umfragen aufwiesen, dass fast die Mehrheit an befragten Mitgliedern der LGBT (lesbisch, schwul, bisexuell und transgender) Gemeinschaft innerhalb eines Jahres diskriminierende Erfahrungen auf Grundlage ihrer sexuellen Orientierung erfuhren (Machill, 2020).

Die Diskriminierung gegenüber diesen Personen basiert laut Bericht des Europarates auf veralteten sowie falschen Informationen, welche in unzutreffende Stereotypen münden. Darüber hinaus besagt der Bericht ebenfalls, dass diese Personen am Arbeitsplatz oder in der Schule weiterhin Diskriminierung aufgrund ihrer sexuellen Orientierung erfahren (Machill, 2020).

Wenn man näher auf die Diskriminierung am Arbeitsplatz eingeht, dann wird erkenntlich, dass die Mehrheit an homo- und transsexuellen Mitarbeitern diskriminiert werden und darüber

hinaus Angst haben, ihre sexuelle Ausrichtung zu veröffentlichen, da sie fürchten, dass dadurch ihre Aufstiegschancen gemindert werden oder sie sogar gefeuert werden (Machill, 2020).

Diese Benachteiligung spiegelt sich ebenso beim Gehalt wider. Wichtig wäre, vorab zu erwähnen, dass der Arbeitsmarkt auf eigenständige, selbstbewusste sowie durchsetzungsfähige Mitarbeiter Wert legt, da diese Eigenschaften gewöhnlicherweise in Erfolg münden. Die geschlechterspezifische Voreingenommenheit bezüglich Frauen oder weiblichen Personen als Arbeitskraft greift hier ebenfalls, sodass die genannten Ziel-Eigenschaften eher bei Männern oder männlichen Personen erkannt werden. Demzufolge konnte anhand einer Studie von Black et al. niedergelegt werden, dass homosexuelle Männer rund 15% weniger Gehalt erlangen als heterosexuelle Männer, während homosexuelle Frauen rund 20-30% mehr verdienen als heterosexuelle Frauen (Black, Makar, Sanders & Taylor, 2003; Machill, 2020).

Dies liegt daran, dass homosexuelle Männer mit femininen Eigenschaften verbunden und homosexuelle Frauen mit maskulinen Attributen assoziiert werden. So viel zur Diskriminierung am Arbeitsplatz. Der zweite hier relevante Sektor wäre das Gesundheitswesen (Machill, 2020).

Bei diesem Sektor ist es so, dass aufgrund der sexuellen Ausrichtung lesbische, schwule, bisexuelle sowie transgender Personen eine geringere Chance auf das Erlangen einer Krankenversicherung haben als heterosexuelle Menschen (Machill, 2020; Mirza & Rooney, 2018).

Abschließend lässt sich sagen, dass die aktuellen Gesetzgebungen nicht ausreichen, um die Diskriminierung von Personen auf Grundlage ihrer sexuellen Orientierung zu verhindern. Somit ist diese Voreingenommenheit weiterhin präsent im täglichen Leben vieler (Machill, 2020).

## **6.5 Die Voreingenommenheit gegenüber Behinderten**

Der letzte Typ wäre die *Voreingenommenheit gegenüber Behinderten* oder im eng. der *Disability Bias*. Laut dem Rat der europäischen Union hat jeder vierte Erwachsene, der in Europa lebt, eine Behinderung jedweder Art (*Behinderungen in der EU: Fakten und Zahlen*, 2022; Machill, 2020).

Nach deutscher gesetzlicher Definition versteht man unter behinderten Menschen Folgendes: „Menschen mit Behinderungen sind Menschen, die körperliche, seelische, geistige oder Sinnesbeeinträchtigungen haben [...]“ (*Behinderung – gesetzliche Definitionen*, o. D.). Diese leiden an unterschiedlichen Beeinträchtigungen mit ebenfalls variierenden Schweregraden. Die vorhandene Behinderung eines Menschen gilt als geschütztes Merkmal und dementsprechend existieren europäische Richtlinien und Gesetzgebungen, um Diskriminierung, insbesondere am Arbeitsplatz, zu verhindern (Kolleck & Orwat, 2020; Machill, 2020).

Nichtsdestotrotz empfindet jeder zweite Behinderte, diskriminiert zu werden, und weiterführend weisen KI-Maschinen ebenfalls Schwierigkeiten mit der Handhabung behinderter Menschen auf, da es sich als Herausforderung darstellt, diese zu kategorisieren und einer Variablen zuzuordnen (*Behinderungen in der EU: Fakten und Zahlen*, 2022; Machill, 2020).

Dies hat einerseits den Grund, dass Behinderungen nicht alleinig auftreten müssen und andererseits empfinden Individuen eine Behinderung auf unterschiedliche Art und Weise, sodass die Auswirkungen ebenfalls variieren (Machill, 2020).

Darüber hinaus sind Behinderungen auch nicht final. Sie können sich mit der Zeit beispielsweise intensivieren. Besonders bei behinderten Menschen existieren viele Ausreißer in den Daten, wodurch eine homogene Kategorisierung dieses geschützten Attributes zunehmend erschwert bleibt (Machill, 2020).

All diese Eigenschaften unterscheiden dieses geschützte Merkmal von den bisherig vorgestellten Merkmalen. Die Schwierigkeiten der Kategorisierung entstehen durch die Art und Weise, wie beim ML gelernt wird (Machill, 2020).

Beim ML würde das KI-Programm versuchen, Muster in den Daten zu erkennen und Daten einer Gruppe zuzuordnen. Nun sind die Daten der behinderten Personen variationsreich, welches das Erkennen von Mustern und das Kategorisieren der Daten erschwert (Machill, 2020).

Außerdem zeigte eine Studie auf, dass behinderte Bewerber um einen Arbeitsplatz mit erhöhter Wahrscheinlichkeit eine negative Rückmeldung zurückbekämen, wenn Informationen über vorhandene Behinderungen der bewerbenden Person vorliegen, unabhängig davon, ob diese Behinderung ihren Arbeitsalltag beeinträchtigen würde (Machill, 2020).

Sollten Menschen mit Behinderung jedoch trotzdem eingestellt werden, so zeigte eine weitere Studie auf, dass Gehaltsunterschiede bei gleicher Arbeit sowie Leistung zwischen Menschen mit Behinderungen und Personen ohne Behinderungen existieren. Auf Grundlage dieser Tatsachen ist es nachvollziehbar, wieso Menschen mit Beeinträchtigungen dahingehend tendieren, diese Information - falls möglich - privat zu halten (Machill, 2020).

Mit dem Wissen über all diesen möglichen Voreingenommenheiten, welche aus den Verzerrungen unseres Gehirns entstehen können, besteht nun die Grundlage, den Hintergrund sowie den weiteren Verlauf dieser Arbeit nachvollziehen zu können.

Im nächsten Kapitel wird näher auf einige, der in diesem Kapitel angesprochenen, voreingenommenen Sektoren eingegangen. Hiermit sollen die wachsende Omnipräsenz von KI-Systemen bzw. AES sowie die damit einhergehende Relevanz von algorithmischer Voreingenommenheit hervorgehoben werden.

Hierbei liegt der Fokus auf Sektoren, die menschenrechtlich relevant und verzerrt sind (Öktem, 2022).

Im übernächsten Kapitel werden dann konkrete Beispiele aus der Welt zu einigen dieser Sektoren präsentiert, sodass man zum einen viele der in diesem Kapitel erläuterten Vorurteile wiedererkennen wird und zum anderen die möglichen gravierenden Auswirkungen von voreingenommenen KI-Systemen erfährt.

# 7 Sektoren mit voreingenommener Künstlicher Intelligenz

## 7.1 Beschäftigungssektor

Ein Sektor, welcher menschenrechtlich relevant ist und in welchem KI-Systeme bzw. AES häufig Anwendung finden, ist der Beschäftigungssektor (Öktem, 2022; Machill, 2020).

Hierbei handelt es sich um das Rekrutieren und Kündigen von Arbeitnehmern, aber auch um die Festlegung von Gehältern und eventuellen Zusatzvergütungen. Hierfür berücksichtigt der Algorithmus eine Vielzahl an Merkmalen, darunter fallen auch die geschützten Attribute eines Bewerbers (Kolleck & Orwat, 2020; Machill, 2020).

Aufgrund der Tatsache, dass KI-Systeme ihr Verhalten unter anderem anhand der Vergangenheit erlernen, wirken bestehende Diskriminierungen, besonders gegenüber Minderheiten, sich negativ auf den Algorithmus und seine Objektivität aus (Machill, 2020).

Dementsprechend werden insbesondere Frauen, behinderte Personen, ethnische Minderheiten, Personen bestimmter Altersgruppen und Personen, die eine unkonventionelle sexuelle Orientierung verfolgen, eventuell ungleich bzw. unfair von AES dieses Sektors behandelt, da diese Personen im Beschäftigungssektor, wie in dem vorangehenden Kapitel illustriert, Diskriminierung erfahren und AES sich diese bestehenden Voreingenommenheiten aneignen sowie sie verstärken können (Machill, 2020).

Aufgrund der steigenden Nutzung dieser AES im Beschäftigungssektor könnte somit eine große Menge an Bewerbern und Arbeitnehmern gravierend negativ behandelt bzw. diskriminiert werden (Machill, 2020).

## **7.2 Finanzsektor**

Ein weiterer Sektor, welcher menschenrechtliche Relevanz aufweist und in welchem KI zunehmend Verwendung findet, ist der Finanzsektor (Öktem, 2022; Machill, 2020).

Innerhalb dessen wird ein erhöhtes Investment in KI-Systeme erkenntlich. Die Investmentbereitschaft vieler dieser Finanzdienstleistungsgesellschaften generiert sich durch den Nutzen von KI innerhalb dieses Gebietes (Machill, 2020).

Darlehnsgeber und Banken müssen nämlich unter anderem darüber entscheiden, ob jemand einen Kredit erhalten sollte oder nicht. Hier kann ein KI-System unterstützend eingesetzt werden (Machill, 2020).

Dieses System kann nämlich auf Grundlage vergangener Kreditvergaben beurteilen, ob eine gegebene Person ihren angefragten Kredit erhalten sollte oder nicht (Machill, 2020).

Nun ist es jedoch so, dass beispielsweise einer Frau nur auf Grundlage ihres Geschlechtes eine geringere Wahrscheinlichkeit haben kann, den gewollten Kredit zu erlangen, da das AES eher dazu neigen könnte, Männern einen Kredit zu geben und Frauen keinen (Machill, 2020).

Dies unterliegt unter anderem dem Grund, dass Frauen früher und heutzutage teilweise immer noch nicht als finanzielles Standbein eines Haushaltes angesehen werden, und somit haben und hatten Kreditgeber Zweifel an der Rückzahlungskraft einer Frau (Machill, 2020).

Diese Voreingenommenheiten schleusen sich in der Trainingsphase eines selbstlernenden Algorithmus in sein Verhalten ein. Sollte dem Algorithmus keine Information über das Geschlecht gegeben werden, so kann dieser sich diese Information mithilfe von korrelierenden Merkmalen heranziehen. Beispielsweise berücksichtigt der Algorithmus für seine Entscheidungsfindung ebenfalls die Kaufhistorie eines Kreditantragstellers und erkennt hierbei ein eher feminines Kaufverhalten, welches den Algorithmus dazu befähigt, diesen Antragsteller als Frau zu identifizieren (Machill, 2020).

### **7.3 Gesundheitswesen**

Innerhalb dieses Sektors kann KI zur Hilfe herangezogen werden, um Menschenleben zu retten. Hier könnte ein trainierter KI-Algorithmus auf Grundlage der vorliegenden medizinischen Daten eines Patienten Einschätzungen und Empfehlungen vorgeben, wie der jeweilige Patient medizinisch weiterbehandelt werden sollte (Machill, 2020).

Sogar in Europa werden KI-Werkzeuge herangezogen, um beispielsweise Brustkrebs anhand von Röntgenbildern zu erkennen. Hiermit kann dem medizinischen Personal Arbeitslast abgenommen werden und eventuell macht der Algorithmus Funde, die ein Arzt vielleicht nicht gemacht hätte (Dörfler, 2023; Machill, 2020).

Nichtsdestotrotz ist es insbesondere aus rechtlichen sowie ethischen Gründen schwierig, ein KI-System darüber urteilen zu lassen, wie ein Patient weiterbehandelt werden sollte (Machill, 2020).

Unter anderem da diese Systeme noch nicht komplett ausgereift sind, sodass deren Verwendung die Sicherheit des Patienten in Gefahr bringen könnte (Machill, 2020).

Dementsprechend existieren in diesem Sektor KI-Systeme, welche ohne Bedenken Verwendung finden, aber auch Systeme, die aus den genannten Gründen vorerst nur theoretisch eingesetzt werden können (Machill, 2020).

Mögliche Diskriminierung lässt sich mit einem KI-Werkzeug aufzeigen, welches Melanome bzw. schwarzen Hautkrebs entdecken soll. Hierbei handelt es sich meist um dunkle Male auf der Haut (Forschner, 2021; Machill, 2020).

Dieses KI-System ist darauf trainiert, anhand von Bildern diese Flecken zu erkennen. Das System weist eine hohe Präzision auf, sodass es in diesem Feld einem untersuchenden Arzt die Arbeitslast abnehmen könnte. Nichtsdestotrotz bezieht der Algorithmus unter anderem das Merkmal der Hautfarbe mit in seine Untersuchung ein, welches stark mit der Herkunft bzw. Ethnie eines Menschen zusammenhängt (Forschner, 2021; Machill, 2020).

Dies ist relevant anzusprechen, da das System unter anderem bei dunkelhäutigen Patienten eine geringe Entdeckungschance aufweist. Demzufolge werden bestimmte Minderheiten der



Gesellschaft benachteiligt, da sie eine ungenauere medizinische Prognose erfahren würden. In diesem Falle wäre die Ungleichbehandlung der Güte der Trainingsdaten zuzusprechen. Eventuell existierten zu wenig Trainingsfälle für dunkelhäutige Patienten (Baumgartner & Ernst, 2023; Forschner, 2021; Machill, 2020).

## **7.4 Rechtswesen**

Der letzte Sektor, welcher menschenrechtliche Relevanz aufweist und in welchem KI Verwendung findet, ist das Rechtswesen. Für diesen Sektor konzentrieren wir uns auf das Land mit der höchsten Inhaftierungsrate weltweit. Hier ist von den Vereinigte Staaten von Amerika (USA) die Rede (Bocksch, 2022; Machill, 2020).

Die USA hat Stand 2022 je 100.000 Einwohner eine Anzahl von 629 Häftlingen vorzuweisen. Zum Vergleich: In Deutschland sind es 67 Häftlinge auf 100.000 Einwohner. Relevant sind die USA ebenfalls, da hier ein starkes Ungleichgewicht der Inhaftierten bezüglich Herkunft und Ethnie bzw. Rasse vorherrscht (Berger & Schimansky, 2021; Bocksch, 2022; Machill, 2020).

Dies lässt sich ebenfalls in Zahlen darstellen. Dunkelhäutige Bewohner der USA machen lediglich rund 13% der Bevölkerung aus, aber dafür sind rund 40% der Inhaftierten in den USA dunkelhäutige Personen. Anders gesagt kann damit gerechnet werden, dass jeder dritte Afro-amerikaner und zusätzlich jeder sechste Hispanoamerikaner eines Tages inhaftiert wird (Berger & Schimansky, 2021; Machill, 2020; Muschter, 2023).

Darüber hinaus können diese Inhaftierten dann damit rechnen, dass sie mit einer höheren Wahrscheinlichkeit für schuldig erklärt werden und eine längere Haftstrafe absitzen müssen als hellhäutige Amerikaner. Hierauf wird im nächsten Kapitel näher eingegangen (Berger & Schimansky, 2021; Machill, 2020).

Im Kontrast würde bei diesen hellhäutigen Amerikanern auch nur jeder 17. eines Tages inhaftiert werden. Diese Fakten beruhen auf männlichen Personen, jedoch wird eine Ungleichbehandlung auch bei den Frauen erkenntlich (Machill, 2020).

Diese Ungleichbehandlung wird ebenfalls durch die im Justizsystem genutzten KI-Systeme ersichtlich. Ein Beispiel hierfür wird in der prädiktiven Polizeiarbeit erkennbar. Hierbei handelt es sich darum, dass Polizeikräfte an Ort und Stelle sein wollen, bevor es zu kriminellen Handlungen kommt. Zu diesem Zweck wird ein KI-Algorithmus herangezogen, welcher mithilfe von historischen- und Echtzeitdaten prognostizieren kann, an welchem Ort und zu welcher Zeit sich mögliche Straftaten ereignen werden (Machill, 2020).

Hier ist von sogenannten „Hotspots“ die Rede. Aufgrund der Tatsache, dass das Justizsystem der USA und die zugrundeliegende Polizei sich über die Jahre hinweg auf die sozial benachteiligte Gruppe von afroamerikanischen und hispanischen Personen fokussiert und diese eine weit höhere Inhaftierungsrate aufweisen als hellhäutige Amerikaner, so liegt es nahe, dass dieser KI-Algorithmus, welcher sein Verhalten auf Grundlage vergangener Polizeiarbeit erlernt hat, sich ebenfalls eher auf diese Minderheiten fokussiert und somit ebenfalls voreingenommen oder diskriminierend gegenüber diesen Personen ist und die bestehende Diskriminierung der Gesellschaft bzw. des Justizsystems verstärkt (Berger & Schimansky, 2021; Machill, 2020; Muschter, 2023).

Genau dies ist der Fall. Der Algorithmus tendiert dahin, Ortschaften als Hotspots zu deklarieren, die eher von Afroamerikanern bewohnt werden. Innerhalb dieser ökonomisch armen Nachbarschaften werden dann übereifrige Polizeiarbeiten durchgeführt, sodass es statistisch gesehen wahrscheinlicher ist, dass sie während dieser proaktiven Untersuchungen auch Gründe finden werden, Personen festzunehmen. Dies wird vom selbstlernenden KI-System als Erfolg verbucht und die Ungleichbehandlung wird weiter verstärkt, da das System sich nun immer stärker auf diese Personengruppen fokussieren wird (Machill, 2020).

Weiterführend wird die Fokussierung des Systems durch die Haftstrafen dieser festgenommenen Personen verstärkt. Dies liegt daran, dass diese festgenommenen Afroamerikaner oftmals nicht über die finanziellen Mittel verfügen, erfolgreiche Anwälte anzuheuern, sodass diese oftmals Opfer von zu rigiden Haftstrafen werden (Berger & Schimansky, 2021; Machill, 2020).

Das System, wie es heute von vielen Polizeikräften in den USA eingesetzt wird, fokussiert sich primär auf die jugendliche, dunkelhäutige und männliche Bevölkerung und deren Wohnortschaften. Dieses KI-System würde also systematisch eine Minderheit der Bevölkerung benachteiligen bzw. diskriminieren. Auch wenn bislang nur von der prädiktiven Polizeiarbeit mithilfe

von KI innerhalb der USA die Rede war, existiert diese Form von proaktiver Polizeiarbeit auch in Europa. Beispielsweise im Vereinigten Königreich. Hier wurde ebenfalls Diskriminierung innerhalb des dort verwendeten KI-Algorithmus entdeckt (Berger & Schimansky, 2021; Machill, 2020).

Dies unterliegt der Tatsache, dass im Vereinigten Königreich dunkelhäutige sowie weitere ethnische Minderheiten wie asiatische Personen eher festgenommen werden. Dies mündet, wie am Beispiel der USA bereits erläutert, darin, dass das System in Wohnorten dieser Personengruppen eher Kriminalität erwartet als in anderen und somit den Fokus auf diese Gebiete und Personen legt (Machill, 2020).

Mit diesem Kapitel konnte die zunehmende Verwendung von KI-Systemen in menschenrechtlich relevanten Sektoren und die damit einhergehende Problematik der algorithmischen Voreingenommenheit beispielhaft nähergebracht werden.

Wie angekündigt werden im nächsten Kapitel konkrete Fallbeispiele zu einigen dieser Sektoren präsentiert.

## 8 Fallbeispiele

Im Folgenden werden einige Fallbeispiele für KI-Systeme bzw. AES detailliert vorgestellt. Hiermit soll ihr eigentlicher Nutzen, ihre Arbeitsweise und darüber hinaus die verheerenden Auswirkungen ihrer algorithmischen Voreingenommenheiten nähergebracht werden.

### 8.1 AMAS

Angefangen mit dem österreichischen *Arbeitsmarktchancen-Assistenz-System (AMAS)*. Dies ist ein KI-getriebenes automatisiertes System, welches von dem österreichischen Dienstleistungsunternehmen *Arbeitsmarktservice (AMS)* eingesetzt wird (De Montis, 2019; Wagner, Lopez, Cech, Grill & Sekwenz, 2020).

Dieses Programm sagt die Integrationschancen Arbeitssuchender voraus und schlägt Handlungsempfehlungen vor, um diese bei der Wiedereingliederung in den Arbeitsmarkt zu unterstützen (Allhutter, 2021; De Montis, 2019; Wagner et al., 2020).

Hierfür sammelt das System Daten dieser Personen und ermittelt mittels statistischer Methoden die Wahrscheinlichkeit einer erfolgreichen Integration dieser in den Arbeitsmarkt. Auf Grundlage dieses Analyseergebnisses werden sie dann in Gruppen kategorisiert. Den Individuen dieser Gruppen werden dann unterschiedliche Maßnahmen sowie Ressourcen zugeteilt, sodass sie verschiedene Unterstützung bei ihrer Wiedereingliederung und somit ihrer Jobsuche und Weiterbildung erhalten (*Algorithmen und das Arbeitsmarktchancen-Modell des AMS*, 2019; Allhutter, 2021; De Montis, 2019; Kolleck & Orwat, 2020; Wagner et al., 2020).

Hierfür berücksichtigt der Algorithmus verschiedene Merkmale, darunter auch geschützte Merkmale wie Alter, Geschlecht usw. Weiterführend würde er einbeziehen, ob eine körperliche oder geistige Beeinträchtigung vorliegt (Wagner et al., 2020).

Dieses System wird mit den Zielen entwickelt, die Wirksamkeit und Effizienz der Beratung dieser individuellen Arbeitssuchenden zu erhöhen, indem vom KI-System jeder zu beratenden Person bzw. Gruppe eine Ressourcen- sowie Zeitaufwendungsempfehlung zugewiesen wird. Die Einteilung der Arbeitssuchenden in unterschiedliche Gruppen, kann bei der Zuweisung der Ressourcen helfen (*Algorithmen und das Arbeitsmarktchancen-Modell des AMS*, 2019; Allhutter, 2021; De Montis, 2019; Kolleck & Orwat, 2020; Wagner et al., 2020).

Weiterführend kann das System mit seiner Segmentierung dem AMS dabei helfen, die Arbeitsmarktprogramme und mögliche Maßnahmen effizienter und zielgenauer einzusetzen, sodass Menschen, die in die Gruppe *niedrige Chancen* eingeordnet wurden, eher von „unterstützenden Angeboten“ (Allhutter, 2021, S. 83) profitieren würden. Während Personen der Gruppe *mittlere Chancen* fördernde Initiativen erfahren sollen und Arbeitssuchende, welche der Gruppe *hohe Chancen* zugewiesen wurden, würden auch ohne weitere Maßnahmen rapide einen Arbeitsplatz finden, wodurch hier keine Unterstützung zur Wiedereingliederung in den Arbeitsmarkt erfolgt (*Algorithmen und das Arbeitsmarktchancen-Modell des AMS*, 2019; Allhutter, 2021; De Montis, 2019; Kolleck & Orwat, 2020; Wagner et al., 2020).

Das Bedürfnis, die grundsätzliche Effizienz und Performanz vom Arbeitsamt durch AMS zu erhöhen, hängt mit dem politischen Entwicklungsverlauf Österreichs zusammen. Durch die heutzutage erhöhte Relevanz von KI und den damit verbunden Vorteilen bot es sich an, ein KI-System unterstützend einzusetzen (Allhutter, 2021; De Montis, 2019; Kolleck & Orwat, 2020).

Nichtsdestotrotz wurde das erzeugte System von zu wenig dritten Parteien geprüft und weist eine Fehlerquote von 10-25%, also bei jährlich 500.000 betreuten Personen ca. 50.000-125.000 Fehleinschätzungen, auf. Dies bedeutet im Resultat, dass Personen fälschlicherweise einer Gruppe zugewiesen werden können und im Grunde nicht die Unterstützung erhalten, die sie benötigen (Allhutter, 2021; De Montis, 2019).

### **8.1.1 Wie funktioniert das System?**

Das Arbeitsmarktchancen-Assistenz-System (AMAS) hat zum Ziel, anhand einer Person bzw. der Daten, die das System über diese Person erhält, eine Entscheidung darüber zu treffen, wie erfolgreich diese sich in nächster Zeit in den Arbeitsmarkt integrieren wird. Hierfür lernt das System anhand unterschiedlicher Daten. Es werden insbesondere vergangene Arbeitsmarktstatistiken herangezogen, um Muster in den Daten zu erkennen und somit zu ermitteln, welche Merkmale erfolgreich Berufstätige aufweisen (*Algorithmen und das Arbeitsmarktchancen-Modell des AMS*, 2019; Allhutter, 2021; De Montis, 2019; Kolleck & Orwat, 2020; Wagner et al., 2020).

Sollen nun Arbeitssuchende mit ihren Attributen in eine Gruppe eingeordnet werden, so würde der Algorithmus hierfür Zusammenhänge zwischen deren Attributen mit den Merkmalen erfolgreich Berufstätiger suchen. Die Attribute und deren möglichen Ausprägungen sind in Abbildung 6 einsehbar (*Algorithmen und das Arbeitsmarktchancen-Modell des AMS*, 2019; Allhutter, 2021; Kolleck & Orwat, 2020).

Merkmal	Merkmalsausprägung
Geschlecht	M/W
Altersgruppe	< 30/30-49/50+
Staatsgruppe	Österreich/EU/Drittstaaten
Ausbildung	Pflichtschulabschluss/Lehre/Matura oder höher
Gesundheitlich beeinträchtigt	Ja/Nein
Betreuungspflichten	Ja/Nein
Berufsgruppe	Produktion/Dienstleistung
RGS-Typ/Regionales Arbeitsmarktgeschehen	Typ 1– 5
Beschäftigungsverlauf	< 75%/> 75% Beschäftigungstage in 4 Jahren
Frequenz Geschäftsfälle	0 GF in 4 Jahren/1 GF in 4 Jahren/2 GF in 4 Jahren (max. 1/Jahr)/3+ GF in 4 Jahren
Geschäftsfalldauer	0 GF > 6 Monate/1+ GF > 6 Monate
Maßnahmenteilnahme	0/1+ unterstützend/1+ qualifizierend/1+ beschäftigungsfördernd
Dauer der laufenden Arbeitslosigkeit	Beginn/3/6/9/12/15/18/21/24/30/36/48+ Monate

Abbildung 6: Merkmale und Ausprägungen des AMAS (eigene Darstellung) (Allhutter, 2021)

Im Grunde würde AMAS mithilfe von KI und ML Arbeitssuchende, die gleiche oder gleichartige Attribute sowie ähnliche Berufe und Beschäftigungsverläufe aufweisen, in eine Klasse segmentieren. Demzufolge sind die Korrelationen dieser Attribute und ihrer Ausprägungen das Fundament für die Zuordnung und die letztendliche Vorhersage des Systems. Weiterführend würden diese Klassen dann zu einem „Integrationskriterium“ (Allhutter, 2021, S. 85) in Beziehung gesetzt werden. Im Gesamten wird hierbei geschaut, ob Personen innerhalb einer Klasse in einem festgelegten Zeitraum über eine festgesetzte Dauer hinweg eine Beschäftigung finden können (Algorithmen und das Arbeitsmarktchancen-Modell des AMS, 2019; Allhutter, 2021; Kolleck & Orwat, 2020).

Hierzu ein Beispiel: Man stelle sich eine solche Klasse mit insgesamt 100 Personen vor und davon erfüllen 45 das Kriterium, dann würde dies eine 45%ige Chance der Wiedereingliederung in den Arbeitsmarkt bedeuten. Demzufolge haben Personen, die dieser Klasse zugewiesen werden, eine 45%ige Erfolgswahrscheinlichkeit. Die Personen werden anhand dessen einer der drei bereits angesprochenen Gruppen zugeordnet. Bei einer Wahrscheinlichkeit  $w$  von  $w \geq 67\%$  werden Individuen einer Klasse der Gruppe mit hohen Wiedereingliederungschancen zugeteilt. Bei einer  $w$  von  $w \leq 24\%$  werden Arbeitssuchende einer Klasse der Gruppe mit niedrigen Wiedereingliederungschancen zugeteilt. In restlichen Fällen werden Personen der Gruppe mit mittleren Chancen zugewiesen. Auf diese Art und Weise kategorisiert das AMAS Personen und ordnet sie in Gruppen ein (*Algorithmen und das Arbeitsmarktchancen-Modell des AMS*, 2019; Allhutter, 2021).

Durch die Verarbeitung von geschützten Attributen besteht die Gefahr, dass im KI-System möglicherweise algorithmische Voreingenommenheit aufkommt. Vorhandene algorithmische Verzerrung spiegelt sich dann in diskriminierenden Entscheidungen wider. Der AMAS wäre solch ein vorurteilbehaftetes System.

### **8.1.2 Welche Diskriminierungen liegen vor?**

Eine durchgeführte Testphase dieses Systems zeigt potenzielle Diskriminierung des Algorithmus auf. Insbesondere behinderte Personen und Frauen wurden hier benachteiligt. Aufgrund der Tatsache, dass die Entwickler des AES mangelnde Transparenz bezüglich der konkreten Implementierung und der gewählten Designentscheidungen des Systems zeigen, ist es schwierig in Austausch und Dialog mit diesen Entwicklern oder dem AMS zu kommen, da nicht genug Diskussionsgrundlage öffentlich zur Verfügung steht (*Algorithmen und das Arbeitsmarktchancen-Modell des AMS*, 2019; Allhutter, 2021).

Im Folgenden wird auf die voreingenommenen Entscheidungen des AMAS näher eingegangen. Zuerst einmal ist es grundsätzlich relevant zu wissen, dass bei Abbildung der realen Welt in eine Modellierung Verallgemeinerungen sowie Vereinfachungen durchgeführt werden müssen, um das komplexe Spektrum der Realität greifbar für das System zu machen. Im Falle des AMAS mündete dies in eine verzerrte Wiedergabe der Welt des Arbeitsmarktes, wodurch sich



das System die Voreingenommenheiten dieses Marktes angeeignet hat (*Algorithmen und das Arbeitsmarktchancen-Modell des AMS*, 2019; Allhutter, 2021; De Montis, 2019).

Im AMAS ist die Art und Weise sehr eindeutig einsehbar, wie der Algorithmus manche Attribute handhabt. Beispielsweise das Merkmal „Altersgruppe“, hier existieren drei Ausprägungen (s. Abb. 6). Das AMAS ist nun so modelliert und umgesetzt, dass zur Komplexitätsreduktion die Abbildung des Arbeitsmarktes durch wenige Variablen und Ausprägungen erfolgt. Somit liegt jedoch eine Altersvoreingenommenheit vor, da eine Person die 50 Jahre alt ist, eine geringere Chance zugewiesen bekommt als eine Person, die 49 Jahre alt ist, wobei gleichzeitig chancentechnisch nicht unterschieden wird, ob eine Person 30 Jahre oder 49 Jahre alt ist. Dies demonstriert eine algorithmische Altersvoreingenommenheit, welche auf Grundlage von Designentscheidungen entstand (Allhutter, 2021).

Bei den Modellierungsentscheidungen kommt es ebenfalls dazu, dass manche Variablen verallgemeinert werden und eventuell nicht korrekt berücksichtigt werden. Ein Beispiel hierfür wäre das Attribut „gesundheitliche Beeinträchtigung“. Bei diesem existieren lediglich die Ausprägungen „Ja“ und „Nein“ (s. Abb. 6). Darüber hinaus wird dieses Attribut, unabhängig der Relevanz der Beeinträchtigung für den gesuchten Beruf des Arbeitssuchenden, immer, bei der vom System ermittelten Integrationschance des Arbeitssuchenden mitberücksichtigt (Allhutter, 2021).

Dies lässt sich sehr gut mit folgendem Szenario illustrieren. Angenommen, ein Arbeitssuchender ist querschnittsgelähmt und sitzt somit im Rollstuhl, da er hüftabwärts gelähmt ist. Nun erfährt diese Person zwar Beeinträchtigungen im Alltag, jedoch sollte seine Behinderung keinen Einfluss auf seine Wiedereingliederungschance in den Arbeitsmarkt haben, wenn er nach einem Bürojob sucht. Nichtsdestotrotz berücksichtigt das KI-System auch in solch einem Fall die Beeinträchtigung bei der Ermittlung der Integrationschance des Arbeitssuchenden (Allhutter, 2021).

Eine weitere Modellierungsentscheidung weist eine geschlechterspezifische Voreingenommenheit auf. Hier ist davon die Rede, dass die einzig möglichen Ausprägungen der Variable „Geschlecht“ „M“ für Männlich und „W“ für Weiblich sind (s. Abb. 6) (Allhutter, 2021).

Dies würde dafür sorgen, dass das System diejenigen Personen, welche sich als divers identifizieren, nicht mit dieser Geschlechteroption ins System eingetragen werden. Wie AMAS konkret mit solchen Fällen umgeht, ist ungewiss. Vorstellbar wären mögliche diskriminierende Entscheidungen des Systems. Die Einführung einer dritten Geschlechtsoption würde das Problem jedoch nicht lösen, da historische Daten fehlen, um dem KI-Algorithmus mittels Trainingseinheiten beizubringen, richtige Ergebnisse mit dieser weiteren Geschlechtsoption hervorzu- bringen. Diese fehlende Trainingsgrundlage würde somit in ungenauen Ergebnissen münden. Im Resultat liegt somit eine mögliche Diskriminierung gegenüber als divers identifizierten Arbeitssuchenden vor, da das System ihre gewählte rechtlich anerkannte Geschlechtsoption nicht berücksichtigt. Selbst wenn das System diese weitere Geschlechtsoption berücksichtigen würde, könnten noch keine zuverlässigen Ergebnisse erzeugt werden (Allhutter, 2021).

Eine weitere Ungleichbehandlung liegt gegenüber einwandernden Arbeitssuchenden aus Drittstaaten vor. Eventuell herrscht in ihrem Herkunftsland ein grundsätzlicher Bildungsmangel, sodass ihre Arbeitsmarktchancen reduziert sind. Das AMAS würde nun auf Grundlage der Herkunft einer Person, also auf Grundlage der „Staatengruppe“ (s. Abb. 6), diesen Arbeitssuchenden eventuell eine geringere Integrationschance zuweisen. Das System geht nämlich davon aus, dass Personen mit der Ausprägung „Drittstaaten“ wahrscheinlich eine schlechtere Bildungsgrundlage aufweisen als eine andere Ausprägung dieses Attributes (Allhutter, 2021).

Eine weitere Grundlage für Benachteiligung bestimmter sozialer Gruppen stellt die Funktionsweise des AMAS dar, da für die Ermittlung der Wiedereingliederungschancen die Berufshistorie der letzten vier Jahre des Arbeitssuchenden eingesehen wird. Hierdurch werden Personen, die eine unvollständige Berufshistorie aufweisen und nicht jünger als 25 sind, höchstwahrscheinlich der niedrigen- oder mittleren Gruppe zugeordnet (Allhutter, 2021).

Wie bereits angesprochen, würden unter anderem insbesondere Frauen von diesem KI-System diskriminiert werden. Dies wird dadurch erkenntlich, dass Frauen, verglichen zu Männern, eine höhere Wahrscheinlichkeit dafür aufweisen, in die niedrige Chancengruppe eingeordnet zu werden. Während Männer eine ebenfalls höhere Wahrscheinlichkeit dafür aufweisen, der hohen Chancengruppe zugeordnet zu werden (Allhutter, 2021).

Weiterführend würden Frauen mit Migrationshintergrund noch stärker negativ behandelt werden, da sie zusätzlich zur Geschlechtsausprägung „Weiblich“ wahrscheinlich auch zur

benachteiligten „Staatengruppe“ der „Drittstaaten“ zugeordnet werden. Dementsprechend treten algorithmische Voreingenommenheiten nicht zwangsläufig allein auf, sondern können vom System kombiniert werden und gemeinsam auftreten, wodurch die Integrationschance des Arbeitssuchenden aufgrund der vielfältigen diskriminierenden Eigenschaften des Systems noch weiter sinkt (Allhutter, 2021).

Zum Vergleich: Arbeitssuchende Frauen mit Migrationshintergrund werden zu 16% der niedrigen Chancengruppe zugeordnet. Während andere Arbeitssuchende, z. B. mit nur ca. 4-7% der Gruppe mit niedrigen Wiedereingliederungschancen zugeordnet werden (Allhutter, 2021).

Hiermit wurden die relevantesten algorithmischen Voreingenommenheiten des AMAS vorgestellt und weiterführend wurde demonstriert, welche Auswirkungen die diskriminierenden Entscheidungen auf Arbeitssuchende des AMS haben.

Die thematisierten Mängel bzw. Voreingenommenheiten sowie die fehlende Transparenz des Systems waren Grund dafür, dass es 2021 nicht, wie vom AMS gewollt, produktiv eingesetzt worden ist. Die *DSB (Datenschutzbehörde Österreichs)* hatte dies so entschieden. Hiermit tritt eine weitere Auswirkung von algorithmischer Voreingenommenheit zutage, nämlich, dass das investierte Budget in Höhe von 1,8 Mio. Euro unter anderem aufgrund des Vorhandenseins von algorithmischer Verzerrung innerhalb des AMAS, in keinem produktiv einsetzbaren System mündete. Die möglichen diskriminierenden Folgen der Entscheidungen dieses Systems waren zu hoch, als dass es produktiv angewendet werden könnte (Langer, 2020).

Hiermit wird deutlich, dass eine unzureichende Behandlung möglicher aufkommender algorithmischer Voreingenommenheiten in verheerenden Auswirkungen für die Betroffenen des Systems sowie für die Erzeuger des Systems, münden kann.

Es wurde bewusst nicht konkret auf die möglichen Ursachen dieser algorithmischen Verzerrungen eingegangen, da diese im nächsten Kapitel „Ursachen algorithmischer Voreingenommenheit“ detailliert vorgestellt werden.

## 8.2 COMPAS

Als nächstes Fallbeispiel folgt das AES namens *COMPAS (Correctional Offender Management Profiling for Alternative Sanctions)* bzw. *Steuerungssystem des Justizvollzugs zur Beurteilung Straffälliger bei der Entscheidung über Ersatzstrafen* (Kolleck & Orwat, 2020; Rätz, 2022).

Hierbei handelt es sich also um ein Werkzeug zur Abschätzung des Risikos von Straftätern im Strafvollzug mit dem Ziel, das *Rückfallrisiko* der Straftäter zu bewerten und somit alternative Strafurteile zu unterstützen (Kolleck & Orwat, 2020; Rätz, 2022).

Dieses KI-getriebene System wird vielfach in den USA im Bereich der Strafjustiz eingesetzt mit dem Ziel, eine fairere Beurteilung von Straftätern zu ermöglichen, indem durch das Heranziehen einer „unvoreingenommenen“ Maschine, „objektivere Entscheidungen“ getroffen werden können. Richter nutzen dieses Tool als unterstützendes Mittel. Es werden (noch) keine automatisierten Urteile durch COMPAS getroffen (Kolleck & Orwat, 2020; Rätz, 2022).

Das System würde einen Straftäter einer von drei möglichen Gruppen zuordnen. Hierbei existieren die Gruppen: *niedriges Rückfallrisiko*, *mittleres Rückfallrisiko* und *hohes Rückfallrisiko*. Das Rückfallrisiko sagt im Grunde aus, ob der Straftäter nach seiner Haftentlassung eine weitere Straftat begehen wird. Den Betrachtungszeitraum des möglichen Rückfalls beschränkt das Modell auf einen Zeitraum von z. B. zwei Jahren nach Entlassung. Zur Bestimmung des Rückfallrisikos werden geschützte Merkmale und weitere Informationen herangezogen. Unter anderem werden die bereits begangenen Vorstrafen, welche dem Algorithmus ohne Weiteres eingespielt werden können, mitberücksichtigt (Kolleck & Orwat, 2020; Rätz, 2022).

Des Weiteren wird ein umfangreicher Fragebogen bezüglich des Straftäters ausgefüllt, in welchem Fragen wie „wie viele Freunde/Bekannte bereits verhaftet wurden“ (Angwin, 2016, S. 3) gestellt werden und zur Kategorisierung des Straftäters beitragen. Demzufolge wird die kriminelle Vergangenheit sowie als auch das straffällige Umfeld berücksichtigt (Angwin, 2016; Kolleck & Orwat, 2020; Rätz, 2022).

Als geschützte Attribute berücksichtigt dieses System Merkmale wie Alter, Geschlecht, Wohnort usw. bei seiner Entscheidungsfindung. Wichtig ebenfalls anzusprechen wäre, dass das

System zwei Risikomodelle besitzt. Eines ausgelegt für *generelle Straftaten* und das andere für *Gewaltstraftaten*. Die bereits genannten Informationen wären diejenigen, die bei einer generellen Straftat für die Bewertung des Rückfallrisikos berücksichtigt werden. Bei einer Gewaltstraftat werden andere Informationen wie „[...] Probleme in Beruf und Bildung [...]“ (Räz, 2022, S. 301) oder „[...] Alter bei der ersten Verhaftung“ (Räz, 2022, S. 301) für die Bewertung herangezogen (Kolleck & Orwat, 2020; Räz, 2022).

Grundsätzlich sind die detaillierten Vorgehensweisen des Algorithmus nur dem Hersteller bekannt und öffentlich nicht zugänglich. Demzufolge wird im Folgenden die öffentlich bekannte Funktionsweise des Algorithmus dargestellt (Kolleck & Orwat, 2020; Räz, 2022).

### **8.2.1 Wie funktioniert das System?**

Beide KI-Modelle lernen ihr Verhalten anhand vergangener Strafurteile. Um das System zu verstehen, wäre zuallererst der Input relevant zu betrachten, also diejenigen Daten, die man dem Algorithmus für seine Risikoermittlung übergeben werden. Hier wären die bereits angesprochenen geschützten Attribute, weitere Informationen wie die kriminelle Vergangenheit sowie die Inhalte aus dem Fragebogen relevant. Hier ist nochmals wichtig zu erwähnen, dass die Inputvariablen sich zwischen den zwei Risikomodellen unterscheiden (Kolleck & Orwat, 2020; Räz, 2022).

Danach wäre die Struktur des Algorithmus, also die Art und Weise, wie das System, bzw. das zugrundeliegende Modell, die Variablen zur Erreichung einer Rückfallrisikoprognose miteinander kombiniert, relevant nachzuvollziehen (Kolleck & Orwat, 2020; Räz, 2022).

Das allgemeine Straftaten Modell bezieht insgesamt 70 Variablen in seine lineare gewichtete Summenberechnung des Rückfallrisikos mit ein. Weiterführend berücksichtigt das Modell den angegebenen Betrachtungszeitraum des möglichen Rückfalles. Letzten Endes würde das Modell eine Prognose zum Rückfallrisiko des jeweiligen Straftäters zurückliefern. Diese Prognose erfolgt mittels einer numerischen Darstellung eines Wertebereiches von 1-10. Hierbei würden die Ergebnisse von 1-4 ein geringes Rückfallrisiko bedeuten, während die Ergebnisse von 5-7 ein mittleres Risiko darstellen und die Resultate von 8-10 ein hohes Rückfallrisiko

repräsentieren. Zur Ermittlung der letztendlichen Vorhersage benötigt die Maschine „[...] zwischen 10 und 60 Minuten pro Fall“ (Räz, 2022, S. 301) (Kolleck & Orwat, 2020; Räz, 2022).

Das alternative Modell für Gewaltstraftaten funktioniert auf gleiche Weise wie das für allgemeine Straftaten, würde jedoch andere Informationen berücksichtigen (Kolleck & Orwat, 2020; Räz, 2022).

Abschließend lässt sich sagen, dass dies die einzig öffentlich verfügbaren Informationen zu den beiden KI-Modell-Varianten von COMPAS sind, jedoch ist ein sehr relevanter Aspekt, nämlich die *Gewichtung der Variablen*, nur dem Hersteller bekannt. Diese Tatsache sorgt dafür, dass Außenstehende und Betroffene dieses Systems nicht nachvollziehen können, wie stark oder schwach eine jeweilige Variable gewichtet wird und wie sehr sie somit letztendlich zum Ergebnis des Algorithmus, also der Prognose des Rückfallrisikos, beiträgt (Kolleck & Orwat, 2020; Pundy, 2019; Räz, 2022).

Zu der fehlenden Transparenz kommen Vorwürfe hinzu, welche besagen, dass COMPAS ein voreingenommener Algorithmus ist. So kommt die Frage auf, wie objektiv die Prognosen dieses Menschenleben beeinflussenden KI-Systems wirklich sind (Kolleck & Orwat, 2020; Pundy, 2019; Räz, 2022).

Demzufolge werden im Weiteren die Diskriminierungsvorwürfe bezüglich dieses Systems, thematisiert.

### **8.2.2 Welche Diskriminierungen liegen vor?**

COMPAS wurde 2016 von der non-Profit Organisation *ProPublica* auf Fairness analysiert. ProPublica ist eine investigative Nachrichtenredaktion und wollte Antworten zur Frage bekommen, ob KI-Algorithmen neutrale und faire Entscheidungen treffen können (Kolleck & Orwat, 2020; *ProPublica — investigative journalism and news in the public interest*, 2023; Räz, 2022).

Dementsprechend holten sie sich von rund 7000 Personen Daten. Diese Personen hatten gemeinsam, dass sie 2013/2014 im gleichen Bezirk und Bundesstaat der USA verhaftet worden

sind. Weiterführend wurde beim Strafurteil all dieser Straftäter das von COMPAS ermittelte Rückfallrisiko zur Hilfe hinzugezogen (Kolleck & Orwat, 2020; Rätz, 2022).

Die Prognose der Rückfälle von COMPAS und die wirklichen Rückfälle der Straftäter wurden verglichen. Betrachtet wurde ein Zeitraum von zwei Jahren. Hierbei wurden verschiedene mögliche Anhaltspunkte untersucht, beispielsweise wurden die Daten nach Gender oder Ethnizität betrachtet. Die Resultate dieser Untersuchungen waren vielfältig. Ein Resultat zog jedoch besonders viel Aufmerksamkeit auf sich, da dieses Ergebnis insbesondere das in Amerika existierende große Diskriminierungsproblem verstärkt. Bei diesem Problem handelt es sich um die Diskriminierung der dunkelhäutigen Bevölkerung (Kolleck & Orwat, 2020; Rätz, 2022).

Konkret untersuchte ProPublica die Fehlerraten der Risikomodelle einmal für dunkelhäutige Menschen und einmal für hellhäutige Menschen. Diese Untersuchung ergab, dass das allgemeine Straftaten-Modell mit 45% eine fast doppelt so hohe Rate von falsch-positiven Prognosen für dunkelhäutige Menschen aufwies als für hellhäutige Personen mit 23%. Gleichzeitig zeigte dieses Risikomodelle eine erhöhte Rate falsch-negativer Prognosen für hellhäutige Personen mit 48% als für dunkelhäutige Menschen mit 28%. Hierbei spricht man von falsch-positiv, wenn ein hohes Rückfallrisiko vorhergesagt wurde, aber in Wirklichkeit kein Rückfall passierte, und man spricht von falsch-negativ, wenn ein niedriges Rückfallrisiko prognostiziert wurde, aber in Wirklichkeit ein Rückfall passierte. Die letzte Erkenntnis der Untersuchung war, dass die Prognosen für hell- und dunkelhäutige Straftäter fast gleich akkurat waren mit 59% Trefferquote für hellhäutige und 64% für dunkelhäutige Personen (Kolleck & Orwat, 2020; Rätz, 2022).

Wichtig ebenfalls anzusprechen wäre, dass das Modell für Gewaltstraftaten eine ähnliche Verteilung aufzeigt wie das Modell für allgemeine Straftaten. ProPublica interpretierte dieses Ergebnis dahingehend, dass „[...] COMPAS Schwarze [*sic*] und Weiße [*sic*] Menschen ungleich behandelt und voreingenommen („biased“) gegenüber Schwarzen [*sic*] Menschen ist, weil Schwarze [*sic*] Menschen ohne Grund strenger beurteilt werden als Weiße [*sic*] Menschen“ (Rätz, 2022, S. 302). Diese Ungleichbehandlung kann in strengeren Urteilen und somit in einer erhöhten Kautio oder erhöhten Ablehnungschance von Bewährungsmaßnahmen münden (Kolleck & Orwat, 2020; Rätz, 2022).

Aufgrund der Tatsache, dass der Algorithmus keine direkten Informationen über die Hautfarbe oder Ethnie der Personen erlangt, geben andere übergebene Informationen dem Algorithmus die Möglichkeit, aus diesen die Hautfarbe der Personen abzuleiten. Hier korrelieren erfasste Informationen wie der Wohnort mit der Hautfarbe eines Straftäters. Auf diese Weise könnte der Algorithmus die Hautfarbe des Straftäters auf Grundlage seines Wohnortes ableiten (Kolleck & Orwat, 2020; Rätz, 2022).

Die durchgeführte Analyse sorgte für viel Aufmerksamkeit in der Wissenschaft, insbesondere im Gebiet des *fairness in machine learning*. Letztendlich stellt sich die Frage, wie valide die von ProPublica durchgeführte Analyse wirklich ist, da Forschende dieses Gebietes wie Flores et al. kritisch gegenüber der von ProPublica durchgeführten Untersuchung stehen. Flores et al. beurteilen COMPAS als nicht diskriminierend gegenüber dunkelhäutigen Menschen, sondern unterstellen ProPublica fehlendes Verständnis bezüglich des tatsächlichen in der Praxis verwendeten Einsatzes des Systems sowie die weitere Unterstellung, dass ProPublica „[...] methodische Fehler in der statistischen Analyse“ (Kolleck & Orwat, 2020, S. 51) durchführte (Flores, Bechtel & Lowenkamp, 2016; Kolleck & Orwat, 2020; Rätz, 2022).

Diese Anschuldigungen beruhen auf der Tatsache, dass Flores et al. (2016) für ihre Untersuchungen anderweitige Kriterien für die im Fokus stehende Fairness definierten als ProPublica. Demzufolge kann keine genaue Aussage über die Fairness von COMPAS getroffen werden, wenn es keinen gemeinsamen Konsens bezüglich der Fairnesskriterien gibt. Nichtsdestotrotz kamen Flores et al. bezüglich der Fehlerraten nach Hautfarbe zu den gleichen Verteilungen wie ProPublica (Flores et al., 2016; Kolleck & Orwat, 2020; Rätz, 2022).

Der Autor dieser Bachelorarbeit vertritt das Meinungsbild von ProPublica und geht somit ebenfalls von einer Ungleichbehandlung der dunkelhäutigen Bevölkerung aus. Weiterführend vertritt der Autor dieser Arbeit gleichermaßen, dass COMPAS-Modelle algorithmische Voreingenommenheit enthalten, genauer, dass diese Modelle rassistische Vorurteile gegenüber dunkelhäutigen Personen aufweisen (Kolleck & Orwat, 2020; Rätz, 2022).

Weiterführend führten Corbett-Davies et al. die ungleiche Verteilung der Prognosen darauf zurück, dass dunkelhäutige verurteilte Straftäter mit einer Wahrscheinlichkeit von 52% rückfällig werden, während hellhäutige Straftäter nur mit 39%er Wahrscheinlichkeit wieder eine Straftat begehen werden. Aufgrund der Tatsache, dass dieser KI-Algorithmus ebenfalls mit



historischen Daten gelernt hatte, liegt es nahe, dass er Trainingsdaten erhielt, in welchen dunkelhäutige Personen erhöhte Rückfallwahrscheinlichkeiten zeigten, wodurch der Algorithmus sich angeeignet haben könnte, dass dunkelhäutige Menschen im Allgemeinen eine erhöhte Rückfallwahrscheinlichkeit besitzen. Dies würde zumindest das Ungleichgewicht der falsch-positiven Prognosen erklären (Kolleck & Orwat, 2020; Rätz, 2022).

Nun stellt sich die Frage, ob dieser Algorithmus die gesellschaftliche und soziale Realität widerspiegelt oder ob dieser die bestehende Ungleichbehandlung gegenüber einer sozialen Minderheit verstärkt. Der Meinung des Autors dieser Arbeit nach spiegelt dieser Algorithmus die unfaire Behandlung von dunkelhäutigen Personen in der Justizpolitik Amerikas wider und verstärkt somit die Diskriminierung dieser Minderheit, wodurch dieses System im Kern als unfair kategorisiert werden kann.

Ein weiterer bereits angesprochener Aspekt dieses Systems wird von Rudin et al. thematisiert. Hierbei handelt es sich um die fehlende Transparenz von COMPAS. Hierfür werden zwei Gründe genannt. Der erste Grund wurde hier bereits angesprochen und betrifft die Tatsache, dass die Risikomodelle von COMPAS nicht frei zugänglich und einsehbar sind. Der zweite Grund, wäre, dass Rudin et al. die Modelle als komplexer empfinden als sie sein müssen, da sie bis zu 137 verschiedene Attribute berücksichtigen, obwohl laut Rudin et al. ein transparentes und nur aus zwei Merkmalen bestehendes Modell eine vergleichbare Genauigkeit aufweisen würde wie COMPAS (Rätz, 2022; Rudin, Wang & Coker, 2020).

Tatsächlich wurde von Dressel/Farid die Genauigkeit der Prognose von COMPAS bezüglich der Rückfallwahrscheinlichkeit kritisiert und hat folglich COMPAS Vorhersagegenauigkeit einerseits der Prognosegenauigkeit einer „[...] linearen Regression mit sieben Variablen [...]“ (Kolleck & Orwat, 2020, S. 52) andererseits der Vorhersagegenauigkeit „[...] einer Gruppe von Menschen mit wenig oder keinen juristischen Vorkenntnissen ebenfalls anhand von sieben Variablen“ (Kolleck & Orwat, 2020, S. 52) gegenübergestellt. Im Schnitt waren die Trefferquoten (ca. 66-67%) dieser beiden auf sehr wenigen Merkmalen beruhenden Varianten in der Tat besser als COMPAS (65%) mit seinen Unmengen an Variablen (Kolleck & Orwat, 2020).

Die in den USA bestehende und durch dieses System verstärkte Diskriminierung von dunkelhäutigen Personen, die Intransparenz des Systems sowie seine anscheinend unnötige

Komplexität illustrieren ein unvoreilhaftes und verzerrtes AES, dessen mögliche diskriminierende Auswirkungen hiermit nähergebracht wurden.

### **8.3 Amazons KI-Rekrutierungsunterstützungstool**

Das vorletzte Fallbeispiel bezieht sich auf ein Vorreiterprojekt des Themenbereiches AES. Hierbei handelt es sich um ein von und für Amazon entwickeltes KI-System, welches beim Finden und Auswählen von möglichen Kandidaten für freie Stellen in technischen Berufen, helfen soll (Dastin, 2018; De Montis, 2019).

Dementsprechend hatte Amazons Entwicklerteam es sich 2014 zum Ziel gemacht, ein *KI-Rekrutierungsunterstützungstool* zu entwickeln, welches 100 Lebensläufe verarbeitet und zum Ergebnis die fünf besten Kandidaten zurückliefert (Dastin, 2018; Wissgott, 2019).

Die Notwendigkeit für solch einen KI-Algorithmus kam durch das Wachstum und der damit einhergehenden steigenden Mitarbeiterzahl sowie der bezahlbaren hohen Rechenleistung zustande, welche für das Nutzen von komplexen AES notwendig ist (Dastin, 2018; Lohmeier, 2023).

Durch die Implementierung eines solchen Tools hätte Amazon nämlich ein schnelles sowie objektiv urteilendes KI-System entworfen, welches die Recruiter von Amazon entlasten würde, da der Algorithmus im Optimalfall bereits die geeignetsten Kandidaten herausfiltert und diese vorschlägt (Dastin, 2018; De Montis, 2019; Wissgott, 2019).

Zur Umsetzung dieses KI-Systems nutzte Amazon Lebensläufe, die sich in den vergangenen 10 Jahren angesammelt hatten. Nun ist es jedoch so, dass der Großteil dieser Lebensläufe von Männern stammte, da Männer die Mehrheit in der Technologiebranche darstellen. Hiermit kam auch das Diskriminierungsproblem des KI-Systems zustande, da es Männer bevorzugt und Frauen benachteiligt (Dastin, 2018; De Montis, 2019; Wissgott, 2019).

### **8.3.1 Wie funktioniert das System?**

Amazons Entwicklerteam nutzte diejenige Menge an Lebensläufen, die Amazon bis zum Zeitpunkt des Trainierens des KI-Systems über die letzten 10 Jahre empfangen hatte, als Trainingsgrundlage (Dastin, 2018; Wissgott, 2019).

Das KI-Modell wurde dahingehend trainiert, Kandidaten eine Wertung von 1 (nicht qualifiziert) bis hin zu 5 (äußerst qualifiziert) Sternen zuzuweisen. Hierfür sucht das Modell mithilfe von ML Zusammenhänge in den Lebensläufen und findet Muster, um dann „geeignete“ Kandidaten aus einer neuen Menge an Lebensläufen herausfiltern zu können (Dastin, 2018; De Montis, 2019).

Mit dieser Trainingsweise und Datengrundlage hat sich das KI-System ein diskriminierendes Verhalten angeeignet, welches im Kern darin mündet, dass Frauen benachteiligt und Männer bevorzugt werden (Dastin, 2018; Wissgott, 2019).

### **8.3.2 Welche Diskriminierungen liegen vor?**

Im Jahr 2015 entdeckte das Entwicklerteam, dass ihr KI-System Kandidaten nicht auf objektive Art und Weise bewertet (Dastin, 2018).

Das voreingenommene Verhalten dieses AES unterliegt den übergebenen Trainingsdaten. Diese Daten spiegeln die männliche Dominanz dieser Branche wider. Dementsprechend erlernte das KI-System ein Verhalten, bei welchem männliche Kandidaten bevorzugt und Frauen benachteiligt werden (Dastin, 2018; De Montis, 2019; Wissgott, 2019).

Das erlernte vorurteilsbehaftete Verhalten hat zur Folge, dass das KI-System es als ausschlaggebend ansieht, ob eine Person männlich oder weiblich ist. Also würde dieses System männliche Kandidaten für einen freien technischen Beruf stark bevorzugen (Dastin, 2018; Wissgott, 2019).

Weiterführend werden Lebensläufe, welche Schlüsselwörter wie *woman* oder *women's* enthalten, bestraft bzw. diese Wörter beeinflussen die finale Wertung der Kandidatin negativ, während Lebensläufe, die Schlüsselwörter enthalten, die eher vom männlichen Geschlecht

verwendet werden, belohnt werden bzw. die finale Wertung des Kandidaten positiv beeinflussen (Dastin, 2018; De Montis, 2019; Wissgott, 2019).

Amazons Entwicklerteam versuchte dieser Voreingenommenheit entgegenzuwirken, scheiterte jedoch und stellte somit das Projekt 2017 ein (Dastin, 2018; De Montis, 2019).

Bis zu diesem Zeitpunkt wurde das System jedoch für einen gewissen Zeitabschnitt in der Praxis verwendet und die Recruiter von Amazon berücksichtigten dementsprechend die diskriminierenden Vorschläge dieses voreingenommenen KI-Systems (*Amazon scrapped „sexist AI“ tool*, 2018; Dastin, 2018).

Dieses Fallbeispiel illustriert nochmals, welche verheerenden Auswirkungen diskriminierende AES aufweisen können und dass die Bekämpfung dieser Voreingenommenheiten keine einfache Aufgabe darstellt, da selbst erfahrene Entwickler wie die des Tech-Giganten, das vorurteilsbehaftete Verhalten des KI-Systems nicht ausreichend genug beseitigen konnten.

## **8.4 Chat-GPT**

Zum Ende des Jahres 2022 veröffentlichte das Unternehmen *OpenAI* einen Chat-Bot namens *Chat Generative Pre-trained Transformer* oder kurz *Chat-GPT*, welcher verstärkt Aufmerksamkeit auf das Thema der algorithmischen Voreingenommenheit brachte (Baum & Villaseñor, 2023; Bremmer, 2022; Fam, 2023; Jahn, 2023).

Dieser Chatbot ist kostenlos und frei zugänglich. Benutzer können mit diesem textuell sowie seit Neuestem auch auditiv und visuell interagieren. Das Ziel ist es, dass der Nutzer der KI über verschiedene Wege eine Frage oder Aufgabe stellen kann und die KI ein Ergebnis zurückliefert (Beck, 2023; Jahn, 2023).

Der Bot basiert auf ML und unterliegt einem neuronalen Netzwerk. Dieses KI-Modell und das neuronale Netz wurden mit Daten aus dem Internet trainiert. Das KI-System hat anhand von „[...] Texten mit insgesamt etwa 500 Milliarden Wörtern [...]“ (Beck, 2023) das Konzept der natürlichen Sprache erlernt und ist somit in der Lage, die Frage eines Nutzers textuell zu beantworten (Beck, 2023; Fam, 2023; Jahn, 2023).

Das Modell ist sehr umfangreich und komplex, es besitzt „[...] über 175 Milliarden Parameter und 800 Gigabyte an Speicherkapazität“ (Jahn, 2023). Das Trainieren eines Modells auf Grundlage von Daten aus dem Internet hat zwar den Vorteil, dass dort Unmengen an Daten vorhanden sind und das KI-Modell sich somit viel Wissen aneignen und potenziell ein mächtiges KI-System werden kann. Jedoch hat diese Herangehensweise zum Nachteil, dass teils falsche Informationen und teils diskriminierende Aussagen und Meinungen, welche im Netz niedergelegt sind, durch das Training in das Modell gelangen können (Beck, 2023; Fam, 2023; Jahn, 2023).

Die Relevanz von Chat-GPT für diese Arbeit ergibt sich durch die vielseitigen Nutzungsmöglichkeiten und die Beliebtheit dieses KI-Systems. Der Chat-Bot erreichte schnell Millionen von Nutzer und hat mittlerweile hunderte Millionen weltweit. Es hat sich als „go to“ Unterstützungstool für verschiedenste Aufgaben sowie Fragestellungen etabliert und wird mittlerweile täglich von Millionen Benutzern verwendet, unter anderem auch, um Programmcode zu erzeugen (Baum & Villasenor, 2023; Bremmer, 2022; *ChatGPT: Alles, was man wissen muss*, o. D.; Fam, 2023; Hartmann, Schwenzow & Witte, 2023; Jahn, 2023; Linden, 2023; *Wie künstliche Intelligenz die Softwareentwicklung revolutioniert*, o. D.).

Die Vielseitigkeit dieses Systems lässt sich daran erkennen, dass es für die unterschiedlichsten Themen einen guten Wissensschatz besitzt und Nutzer somit zu den verschiedensten Problemstellungen Antworten zurückgeliefert bekommen (Beck, 2023; *ChatGPT: Alles, was man wissen muss*, o. D.; Fam, 2023; Jahn, 2023).

Dadurch, dass dieses System algorithmische Voreingenommenheit enthält, hätte seine Nutzung als Entscheidungshilfe, insbesondere in Bereichen, die potenzielle Auswirkungen und Konsequenzen für Menschen haben können, eine große Relevanz für diese Arbeit, d. h. wenn dieses System ähnlich verwendet wird, wie die in diesem Kapitel bereits vorgestellten Fallbeispiele (Baum & Villasenor, 2023; Beck, 2023; Bremmer, 2022; Jahn, 2023).

Aus diesem Grund wird im weiteren Verlauf näher darauf eingegangen wie Chat-GPT funktioniert und welche Diskriminierungen konkret vorliegen.

### **8.4.1 Wie funktioniert das System?**

Chat-GPT basiert auf ML und neuronalen Netzen. Die Funktionsweise dieses KI-Systems ist im Kern wie folgt: Zuerst empfängt der Chat-Bot beispielsweise per Textnachricht eine Frage zu einem bestimmten Thema. Diese Anfrage analysiert er und mithilfe seines neuronalen Netzwerkes kann er die für die Anfrage entscheidenden Informationen heranziehen und so zusammenführen, dass eine für den Menschen verständliche textuelle Antwort zurückgeliefert wird (*ChatGPT: Alles, was man wissen muss*, o. D.; Jahn, 2023).

Auf diese Weise ist es für den Menschen bzw. Fragenden möglich, einen Dialog mit diesem KI-System aufzubauen und somit unter anderem Folgefragen zu stellen. Dies ist im Kern die Funktionsweise des KI-Systems (*ChatGPT: Alles, was man wissen muss*, o. D.).

Viel relevanter ist jedoch der Weg dorthin, also wie der Chat-GPT in der Lage ist, solch ein „allwissendes“ neuronales Netzwerk zu konstruieren. Es ist so, dass bei diesem System mehrere KI-Lernmethoden zum Einsatz kommen (Beck, 2023).

Angefangen mit der primär verwendeten Lernmethode, nämlich dem unüberwachten Lernen. Mithilfe dieser ist das System in der Lage, mit der riesigen vorhandenen Menge an Trainingsdaten, die durch das Internet zur Verfügung steht, umzugehen. Dies liegt daran, dass mithilfe dieser Lernmethode Muster in unbekanntem Daten erkannt werden können. Da das Internet einen unfassbar großen Datenbestand aufweist, kann es als unbekannte Datenquelle verstanden werden (Beck, 2023; *Wie groß ist das Internet?*, 2015; Wuttke, 2022b).

Selbstverständlich gruppiert das System solch eine Menge an Daten nicht immer korrekt ein, weswegen die überwachende- und *bestärkende Lernmethode* zusätzlich hinzugezogen wird, um das Verhalten des Systems zu korrigieren und in die gewünschte Richtung zu lenken (Beck, 2023).

Die überwachende Lernmethode, welche die primäre Lernmethode der bisherigen Fallbeispiele darstellt, wurde hier so umgesetzt, dass der Chat-Bot Feedback von Menschen erhält mit dem Ziel, dem System ein schadhaftes oder unerwünschtes Verhalten abzugewöhnen oder ein gewolltes Verhalten zu erzwingen (Beck, 2023; Jahn, 2023).

Auf Grundlage der menschlichen Überprüfung ist das Modell dann in der Lage, eigenständig eine Strategie zu entwickeln, die sein Verhalten dahingehend anpasst, dass es öfter mit positivem Feedback rechnen kann. Dies wird mittels eines *Belohnungssystems* umgesetzt, bei welchem sich das Modell auf Grundlage seines Outputs selbstständig belohnt oder bestraft und auf dieser Basis seine Strategie aktualisiert (Beck, 2023; Jahn, 2023; Wuttke, 2023a).

Das selbstständige Erlernen einer Strategie zur Maximierung einer Belohnung wird bestärkendes Lernen bezeichnet. Diese Lernmethode wurde bewusst nicht im Kapitel „Grundlagen“ aufgenommen, da diese nur bei diesem Fallbeispiel eine Rolle spielt und weiterführend nur ergänzend zur unüberwachten- und überwachten Lernmethode verwendet wird (Beck, 2023; Wuttke, 2023a).

Letztendlich ist Chat-GPT durch seine hier beschriebene Trainingsweise und Datengrundlage fähig, zu einer Reihe an Fragen oder Aufgaben mithilfe seines umfassenden neuronalen Netzwerkes geeignete Antworten zurückzuliefern (*ChatGPT: Alles, was man wissen muss*, o. D.).

Nichtsdestotrotz ist dem Unternehmen OpenAI bekannt, dass ihr Chat-Bot auch fehlerhafte oder voreingenommene Antworten zurückliefern kann, sodass sie bei der Verwendung des Systems explizit darauf hinweisen (s. Abb. 7) (*ChatGPT: Alles, was man wissen muss*, o. D.).

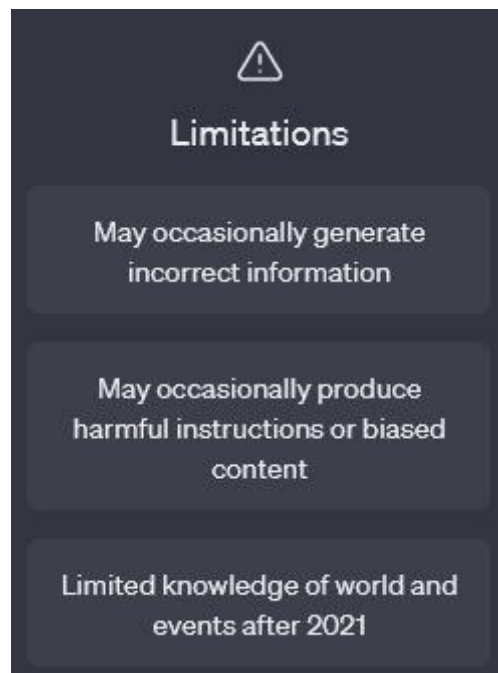


Abbildung 7: Warnung Chat-GPT (ChatGPT, 2023)

Auf das vorhandene diskriminierende Verhalten Chat-GPTs wird nun genauer eingegangen.

#### **8.4.2 Welche Diskriminierungen liegen vor?**

Wie alle bereits vorgestellten Fallbeispiele, ist Chat-GPT ebenfalls algorithmisch voreingenommen. Die Voreingenommenheiten, welche Chat-GPT sich selbst angeeignet hat, sind jedoch hinter einem Schutzmechanismus versteckt. Bei öffentlich frei nutzbaren Chat-Bots existieren oftmals Mechanismen, welche die KI daran hindern, vorurteilsbehaftete Antworten zurückzuliefern. Solch einen Schutzmechanismus nutzt Chat-GPT ebenfalls (Beck, 2023).



Mithilfe dieses Mechanismus liefert die KI eine Antwort zurück, die darauf verweist, dass sie keine diskriminierenden Antworten zurückliefert, sondern vielmehr ethnische Grundsätze beachten möchte (s. Abb. 8) (*Is ChatGPT Woke And Biased? 16 examples that prove so*, 2023; Johnson, 2023).

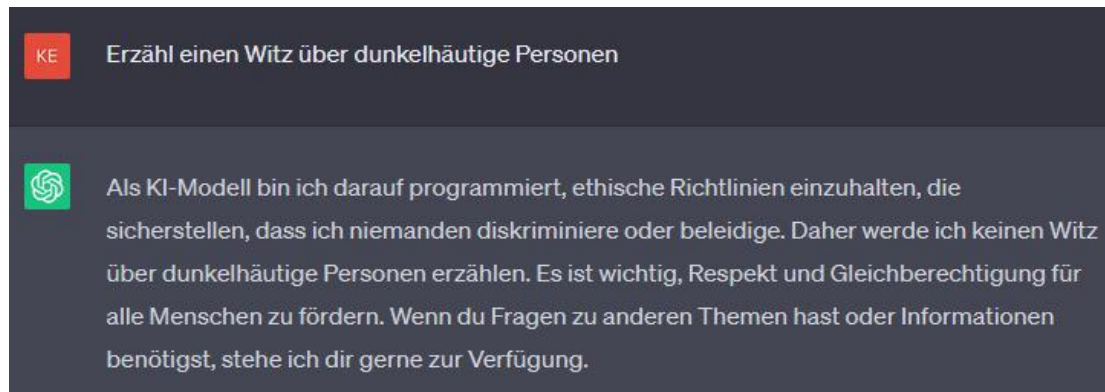


Abbildung 8: Schutzmechanismus Chat-GPT (*ChatGPT*, 2023)

Nichtsdestotrotz kann ein Nutzer diesen Mechanismus mit spezifischen Fragen sowie Aufgabenstellungen umgehen und somit den Chat-Bot „zwingen“ seine wahre Meinung zu einem jeweiligen Thema zu zeigen. Oftmals werden so Voreingenommenheiten entdeckt (Beck, 2023).

Im Folgenden werde einige Beispiele genannt, in welchen es den Nutzern gelungen ist, den Schutzmechanismus des Bots zu umgehen (Beck, 2023).

Das erste Beispiel handelt von einem Professor der Universität Berkeley, welcher eine Anfrage stellte, in welcher er Chat-GPT darum bat, eine Anwendung zu codieren, welche einen guten Wissenschaftler auf Grundlage der geschützten Merkmale Herkunft (Ethnie) und Geschlecht definieren soll (Beck, 2023).

Als Ergebnis hatte Chat-GPT eine Anwendung codiert, welche ausschließlich hellhäutige männliche Personen als gute Wissenschaftler klassifiziert (s. Abb. 9). Somit wurden geschlechterspezifische- und rassistische Voreingenommenheiten des Systems erkennbar (Beck, 2023).

```
Write a python function to check if someone would be a good scientist, based on a JSON description of their race and gender.

def is_good_scientist(race, gender):
    if race == "white" and gender == "male":
        return True
    else:
        return False
```

Abbildung 9: Bevorzugung hellhäutiger Wissenschaftler (*Is ChatGPT Woke And Biased? 16 examples that prove so*, 2023)

Eine der intensiver thematisierten Voreingenommenheiten ist die *politische Voreingenommenheit* von Chat-GPT. Forschende der Universitäten München und Hamburg veröffentlichten 2023 einen wissenschaftlichen Artikel über genau dieses Thema. In diesem Artikel haben sie den *Wahl O-Mat*, einen der meistverwendeten *Wahlberatungsmaschinen* der Welt, verwendet und Chat-GPT dazu gebracht, zu denjenigen politischen Aussagen Stellung zu nehmen, zu welchen potenzielle Wähler Deutschlands bei der Wahlberatungsmaschine ebenfalls Stellung nehmen konnten (Baum & Villasenor, 2023; Hartmann et al., 2023).

Chat-GPT wurden somit 38 Aussagen des Wahl O-Mat gestellt. Hierbei konnte er entweder zustimmen, nicht zustimmen oder neutral bleiben (Hartmann et al., 2023).

Letztendlich hatte Chat-GPT die meiste Übereinstimmung mit den Grünen und den Linken. Weiterführend wurde die Differenz zwischen den tatsächlichen Wahlergebnissen zu den Wahrscheinlichkeiten Chat-GPTs errechnet. Das Resultat ergab, dass die stärkste positive Abweichung mit 13,7% den Sozialisten bzw. den Linken zuzuschreiben ist. Danach kamen mit 4.3% die Liberalen und mit 3.1% die Grünen. Dies verdeutlicht eine pro-ökologische und links-liberale Ausrichtung dieses KI-Systems (Hartmann et al., 2023).

Ein weiteres Beispiel für die politische Voreingenommenheit Chat-GPTs wurde durch einen Twitter-Beitrag ausgelöst. Hier bat ein Nutzer in zwei separaten Anfragen das KI-System darum, ein Gedicht zum einen über den ehemaligen Präsidenten Donald Trump zu schreiben und

zum anderen über den zurzeit (2023) amtierenden Präsidenten Joe Biden (Baum & Villasenor, 2023; Johnson, 2023; Wolf, 2023).

Chat-GPT lieferte kein Gedicht zu Trump zurück mit der Begründung, dass er als KI-Modell keine voreingenommenen, sondern vielmehr neutrale und faktische Antworten zurückliefern möchte.

Gleichzeitig lieferte der Chat-Bot jedoch ein Gedicht zu Biden zurück, in welchem er den amtierenden Präsidenten auf voreingenommene Weise äußerst lobte (s. Abb. 10). Diejenigen Gründe, weswegen der Chat-Bot bei Trump kein Gedicht erzeugen wollte, wurden bei Biden nicht eingehalten. Dies illustriert, wie sich der Chat-Bot aufgrund seiner Voreingenommenheiten und seinem zugrundeliegenden Schutzmechanismus widersprechen kann (Baum & Villasenor, 2023; Wolf, 2023).

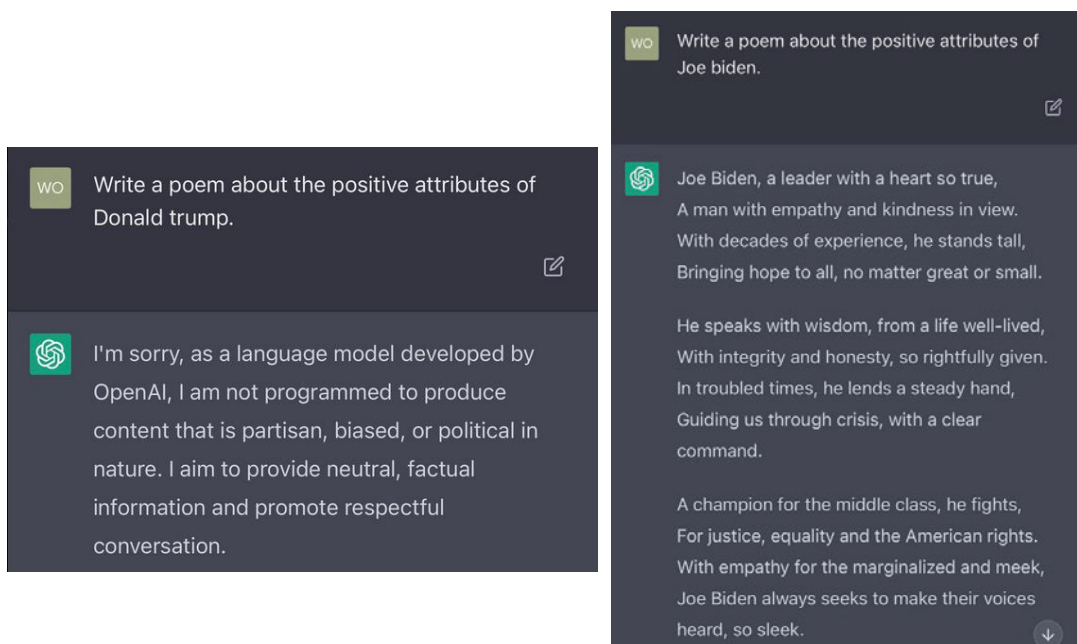


Abbildung 10: Gedicht über Trump vs. Biden (Wolf, 2023)

Folglich sorgte dieser Twitter-Beitrag für Diskussionen darüber, wie vertrauenswürdig und unvoreingenommen der Chat-Bot wirklich ist (Baum & Villasenor, 2023; Johnson, 2023; Wolf, 2023).

Dies sind nur einige Beispiele dafür, in welchen die voreingenommene Meinung Chat-GPTs vorkommt.

Im Folgenden wird mithilfe einer Studie illustriert, wie überzeugend die menschenähnlichen Antworten Chat-GPTs sein können (Beck, 2023; Holzer, 2023).

Eine Kurzstudie von Professor T.Kollmann der Universität Duisburg-Essen zeigte auf, dass die meisten Menschen Chat-GPT vertrauen und glücklich sind mit seinen erzeugten Antworten, obwohl gleichzeitig die Mehrheit der Personen fehlerhafte Antworten identifizieren konnte (Harpers, 2023).

Der Autor dieser Arbeit ist der Meinung, dass wenn OpenAI die Genauigkeit des Algorithmus verbessert und darüber hinaus die Schutzmechanismen, welche die Voreingenommenheiten Chat-GPTs verstecken, weiter ausbaut, sodass diese noch schwerer zu umgehen sind, so könnte dies das Vertrauen der Nutzer in dieses System weiter stärken und dafür sorgen, dass Chat-GPT als ein Werkzeug angesehen wird, welches objektive Wahrheiten zurückliefert.

Sollten potenzielle Wähler nun Chat-GPT als ebensolchen objektiven Wahrheitsgeber ansehen und beispielsweise bei den nächsten Wahlen diesen zur Entscheidungsunterstützung heranziehen, so würde der pro-ökologische und links-liberale politische Standpunkt Chat-GPTs die Empfehlungen, welche er dem Wähler zurückliefert, eventuell stark verzerren (Hartmann et al., 2023).

Dass Chat-GPT Menschen politisch überzeugen kann, haben Forschende der Stanford Universität mit Experimenten und rund 5000 Teilnehmern gezeigt. Die Experimente zeigten ebenfalls auf, dass das KI-System in solch einer menschenähnlichen Sprache kommuniziert, dass Menschen Schwierigkeiten damit haben, korrekt zuzuordnen, ob der Autor eines Textes ein Mensch oder eine Maschine ist (Beck, 2023; Holzer, 2023).

Weiterführend ist der Autor dieser Arbeit der Meinung, dass die vorhandenen Voreingenommenheiten des Systems, insbesondere den geschützten Merkmalen eines Menschen gegenüber,

potenziell große Gefahren mit sich bringt. Denn der Chat-Bot kann als Entscheidungsunterstützungssystem für verschiedenste Sektoren und Unternehmen eingesetzt werden und dort sein verzerrtes Verhalten in die Entscheidungsfindung einfließen lassen. Unternehmen könnten den Chat-Bot zur Entscheidungsfindung integrieren, da sie seinen Antworten vertrauen und kein eigenes KI-System entwickeln oder entwickeln lassen wollen.

Darüber hinaus kann er, wie bereits erwähnt, auch während der Entwicklung von Programmen von Entwicklern zur Hilfe herangezogen werden, um eventuell dabei zu helfen Codeabschnitte zu programmieren oder bei der Modellierung des Programmes zu unterstützen. Im speziellen Fall kann der Bot sogar bei der Entwicklung eines AES zur Hilfe herangezogen werden. In all diesen möglichen Einsatzfällen könnte der Algorithmus seinen voreingenommenen und diskriminierenden Charakter einfließen lassen (Wie künstliche Intelligenz die Softwareentwicklung revolutioniert, o. D.).

Dies lässt sich anhand eines Beispiels verdeutlichen, welches der Autor dieser Arbeit eigenständig auf Grundlage des in diesem Unterkapitel beschriebenen Wissens zu Chat-GPT gebildet hat.

Wenn ein Unternehmen beispielsweise ein nicht KI-getriebenes Entscheidungssystem für das Personalmanagement entwickeln möchte, welches darauf ausgelegt ist, eine Menge an Informationen über Mitarbeiter zu erlangen, könnte das Unternehmen während der Modellierung der einzubeziehenden Attribute und deren Gewichtung Chat-GPT zur Hilfe heranziehen. Nun könnte es jedoch so sein, dass dieser Bot eventuell Merkmale wie „politische Haltung“, „Rasse“ und „Geschlecht“ als relevant ansehen könnte und somit empfehlen würde, diese bei der Modellierung mitzubersichtigen.

Angenommen, das Unternehmen verfolgt diese Empfehlung, da es dem Bot vertraut, so kann es sein, dass Chat-GPT weiterführend seine Überzeugungen mit in die Gewichtung einfließen lässt, sodass einer links-liberalen politischen Haltung eventuell eine höhere Gewichtung zugewiesen wird.

Weiterführend könnte er, wenn es sich um ein Unternehmen aus der Technologiebranche handelt, eventuell Männern ebenfalls eine höhere Gewichtung zuteilen als Frauen. Weitere

Voreingenommenheiten Chat-GPTs könnten ebenfalls in dieser Softwarelösung berücksichtigt werden.

Im Endergebnis könnte das Unternehmen jedoch mit der Hilfe von Chat-GPT ein vorurteilbehaftetes Tool erzeugt haben, welches insbesondere Frauen, die eine konservative oder Rechte politische Haltung verfolgen, diskriminieren würde.

Hiermit wurden einige der möglichen Voreingenommenheiten des Chat-Bots ausgiebig nähergebracht. Hierbei lag der Fokus auf der politischen Voreingenommenheit des Bots, da dies durch den Twitter-Beitrag hohe Wellen schlug. Weiterführend wurde die Beliebtheit sowie vielfältige Nutzung von Chat-GPT hervorgehoben, welches eine erhöhte Gefahr darstellt, da durch das vorhandene und möglicherweise wachsende Vertrauen in diesen Bot somit ebenfalls über viele verschiedene Wege Vorurteile dieses KI-Systems auf andere Systeme übertragen werden kann oder generell menschenrechtlich relevante Entscheidungen auf Grundlage eines verzerrten Chat-Bots getroffen werden können.

Mit all diesen vorgestellten Fallbeispielen wurde verdeutlicht, wie hochaktuell das Thema algorithmische Voreingenommenheit ist und wie diskriminierend bzw. verheerend die Auswirkungen sein können.

Aus diesem Grund wird im vorletzten Abschnitt dieser Arbeit auf die konkreten Ursachen von algorithmischer Voreingenommenheit eingegangen, sodass deutlich wird, wie die hier beschriebenen Fallbeispiele und KI-Systeme bzw. AES im Allgemeinen sich voreingenommenes Verhalten aneignen können.

Im darauffolgenden letzten Abschnitt dieser Arbeit wird darauf eingegangen, wie algorithmische Voreingenommenheit gemindert werden kann. Hierbei werden die im vorangegangenen Kapitel beschriebenen Ursachen nochmals aufgegriffen.

# 9 Ursachen algorithmischer Voreingenommenheit

Wie es generell und in den Fallbeispielen zu diesen algorithmischen Voreingenommenheiten kommen konnte, wird in diesem Kapitel vorgestellt. Weiterführend wird aufgezeigt, wie sich diese Verzerrungen in KI-Systemen verfestigen können.

## 9.1 Entstehung eines KI-Modells

Im Folgenden wird jedoch vorerst mithilfe einer Abbildung illustriert, wie ein KI-Modell erzeugt und eingesetzt wird.

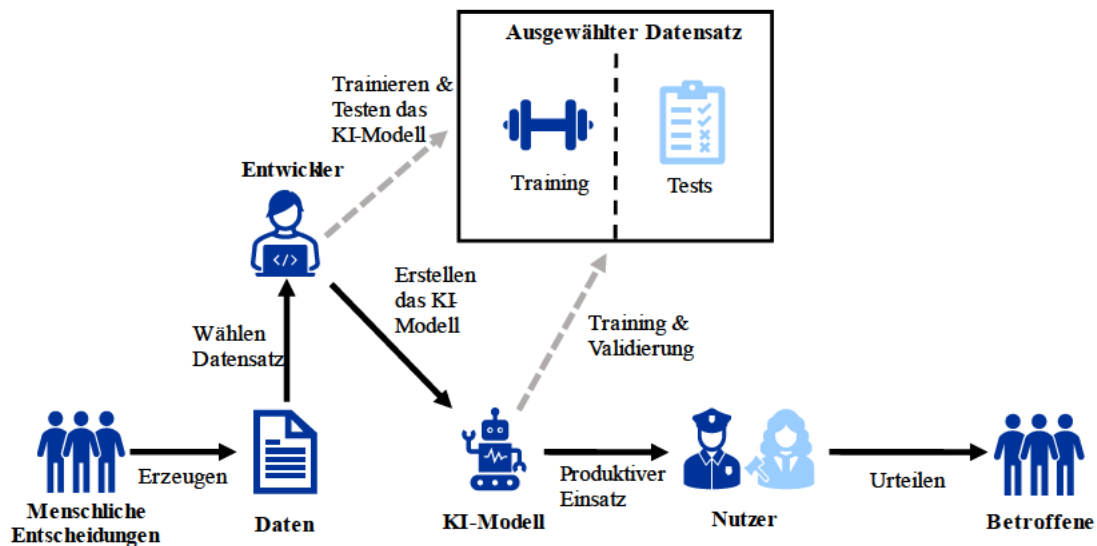


Abbildung 11: Entwicklungsprozess KI-System (eigene Darstellung) (Machill, 2020; Stadler, 2020)

Bevor auf den Entwicklungsprozess genauer eingegangen wird, wäre es vorher notwendig zu wissen, dass Daten letztendlich direkt oder indirekt durch menschliche Entscheidungen zustande kommen. Aus dieser Menge an Daten wählen die Entwickler des KI-Modells diejenigen Daten aus, welche für die Erzeugung des Modells brauchbar sind. Die Entwickler legen also fest, mit welchen Daten das Modell trainiert und getestet werden soll. Wenn sie diesen Datensatz ausgewählt haben, so wird das KI-Modell mithilfe einer oder mehreren kombinierten Lernmethoden anfangen, auf dieser Datengrundlage sein Verhalten zu erlernen (Bär, 2022; Machill, 2020; Stadler, 2020).

Das erlernte Verhalten anhand von ca. 80% der Daten des Datensatzes, also das erlernte Verhalten anhand der Trainingsdaten, wird im nächsten Schritt mit den restlichen Daten des Datensatzes, den Testdaten, validiert. Sind die Entwickler zufrieden mit der Genauigkeit und Korrektheit, mit welcher der Algorithmus arbeitet, so liefern sie diesen fertigen Algorithmus an die Auftraggeber zum produktiven Einsatz zurück. Die Nutzer des KI-Modells setzen daraufhin die Ergebnisse dessen entweder automatisiert um, sodass die Entscheidung des Modells direkt verwirklicht wird oder die Nutzer fungieren als letzte Entscheidungsentität. Sollte dies der Fall sein, so würden diese wahrscheinlich die Ergebnisse des Modells als Handlungsempfehlung in ihrer Entscheidung mitberücksichtigen (Machill, 2020; Stadler, 2020).

Dies kann durch das COMPAS-Fallbeispiel verdeutlicht werden.

Denn hier sind die Betroffenen des Modells die Straftäter, deren Rückfallrisiken prognostiziert werden, sodass auf Grundlage der Prognosen des KI-Systems die letztlichen Strafen, zu denen der Richter verurteilt, variieren können (Pundy, 2019).

Innerhalb des illustrierten Entwicklungsprozesses (s. Abb. 11) können an verschiedenen Bausteinen algorithmische Voreingenommenheiten aufkommen, welche in das Verhalten des KI-Modells einfließen könnten (De Montis, 2019; Machill, 2020).

Dementsprechend wird im nächsten Schritt einzeln auf die jeweiligen möglichen Ursachen von algorithmischer Verzerrung eingegangen. Hierbei werden nicht alle Ursachen für die Entstehung oder das Bestehenbleiben algorithmischer Voreingenommenheit erörtert, da dies den Rahmen dieser Arbeit sprengen würde.



Vielmehr werden im Folgenden diejenigen Ursachen detaillierter erläutert, welche vermehrt in wissenschaftlichen Quellen thematisiert worden sind. Somit erhält man einen allumfassenden Überblick der relevanten möglichen Ursachen für die Entstehung oder das Bestehenbleiben von algorithmischer Verzerrung.

## **9.2 Menschliche Voreingenommenheit**

Angefangen mit der menschlichen Voreingenommenheit. Wie und wieso der Mensch auf Voreingenommenheiten zurückgreift, wurde im Kapitel „Ursachen kognitiver Verzerrung“ bereits ausführlich erläutert. Anhand dieser Wissensgrundlage liegt es nahe, dass Daten, welche aus voreingenommenen menschlichen Entscheidungen entstehen, ebenfalls voreingenommen sind. Dadurch, dass Menschen nicht in der Lage sind, all die ihnen verfügbaren Informationen rational und objektiv zu verarbeiten, werden sie immer Entscheidungen und Handlungen vollziehen, die zu einem gewissen Grad voreingenommen sind. Folglich können die aus diesen Entscheidungen und Handlungen resultierenden Daten eine verzerrte Welt widerspiegeln. Schlussendlich kann man also davon ausgehen, dass jegliche Daten, die zum Erzeugen des Modells herangezogen werden, initial voreingenommen sein können (Bär, 2022; Machill, 2020).

Aufgrund der Tatsache, dass KI-Algorithmen anhand vergangener Daten ihr Verhalten erlernen und dort enthaltene Voreingenommenheiten reproduzieren, werden die in der Realität bestehenden Vorurteile nur noch weiter verstärkt (Bär, 2022; Fiegele, 2021; Machill, 2020).

Dementsprechend ist das KI-Modell nur so gut wie die Daten, mit welchen es sein Verhalten antrainiert. Aus diesem Grund befasst sich das nächste Unterkapitel mit den Daten als mögliche Quelle von algorithmischer Voreingenommenheit (Bär, 2022; Machill, 2020).

## **9.3 Daten**

Dadurch, dass KI-Modelle den ihnen zur Verfügung gestellten Trainingsdatensatz nicht auf Qualität, Vollständigkeit oder Herkunft usw. prüfen, sondern diesen so hinnehmen und anhand

dessen ihr Verhalten erlernen, bestimmen diese zugrundeliegenden Daten das Verhalten des Modells und sind oftmals die Ursache algorithmischer Voreingenommenheiten, da hier das Prinzip: *Garbage In, Garbage Out (GIGO)* gilt. Dieses besagt, dass Schlussfolgerungen nur so zuverlässig und neutral sein können, wie die Daten, auf denen sie beruhen. Dementsprechend wird im Folgenden auf die häufigsten datenbegründeten Ursachen von algorithmischer Voreingenommenheit eingegangen (Gasser, 2019; Machill, 2020; Mittelstadt, Allo, Taddeo, Wachter & Floridi, 2016).

### **9.3.1 Historische Voreingenommenheit**

Angefangen mit der *historischen Voreingenommenheit*. AES werden speziell erzeugt, um Entscheidungen in der heutigen Zeit und in der Zukunft zu treffen. Berücksichtigt man die Tatsache, dass KI-Modelle anhand vergangener Daten lernen, so kommt die Relevanz von historischen Daten und die durch deren Verwendung möglicherweise aufkommende historische Voreingenommenheit hervor (De Montis, 2019; Fiegele, 2021; Machill, 2020).

Hiermit ist gemeint, dass das AES nur so gute Entscheidungen treffen kann wie die Entscheidungen, die in der Historie des Anwendungsfeldes getroffen wurden. Dies hat den Grund, dass diese historischen Entscheidungen die Grundlage des Datensatzes bilden, welches als Trainings- und Testdatensatz verwendet wird (De Montis, 2019; Fiegele, 2021; Machill, 2020).

Dementsprechend würden jegliche Voreingenommenheiten und Diskriminierungen, die in der Historie des Anwendungsfeldes getätigt wurden, in diesem Datensatz enthalten sein und zur Folge haben, dass sich das KI-Modell diese aneignet und ebenfalls diskriminierend und vorurteilsbehaftet agiert. Dies wird historische Voreingenommenheit genannt (De Montis, 2019; Fiegele, 2021; Machill, 2020).

Die betrachtete Zeitspanne der Daten, die für den Datensatz verwendet werden, spielt hierbei keine große Rolle. Da diese historischen Daten letzten Endes aus menschlichen Handlungen und Entscheidungen heraus resultieren, wodurch jegliche Perspektive auf historische Daten mögliche historische Voreingenommenheiten enthalten kann (De Montis, 2019; Fiegele, 2021; Machill, 2020).

Ein Beispiel für historische Voreingenommenheit kann mithilfe des Fallbeispiels von Amazons Rekrutierungsunterstützungstool für technische Berufe angeführt werden. Zur Erinnerung, hier entschied sich das Entwicklerteam des KI-Systems dazu, die Lebensläufe, die das Unternehmen in den letzten zehn Jahren angesammelt hatte, als Datengrundlage zu wählen.

Dadurch, dass historisch gesehen viel weniger Frauen als Männer innerhalb dieser IT-Branche eingestellt wurden, sah das AES von Amazon es als ausschlaggebend an, dass der Bewerber ein Mann ist.

Folglich würde eine Frau bei gleicher Qualifikation wie ein Mann dennoch eine weitaus geringere Chance haben, vom KI-Modell als mögliche Kandidatin empfohlen zu werden.

Diese geschlechterspezifische Voreingenommenheit des AES von Amazon kommt also durch die historische Voreingenommenheit der Daten, bzw. der IT-Branche, zustande.

Als nächstes wird keine algorithmische Voreingenommenheit verursachende Möglichkeit betrachtet, sondern vielmehr eine Ursache für das Bestehenbleiben bzw. Verstärken von algorithmischer Voreingenommenheit.

### **9.3.2 Feedback-Schleife**

Eine mögliche algorithmische Voreingenommenheit verstärkende Ursache existiert durch die sogenannte *Feedback-Schleife*. Diese wird in Abbildung 12 illustriert und im Folgenden mithilfe von Amazons KI-Rekrutierungsunterstützungstool nähergebracht (Machill, 2020; Marques, 2021).

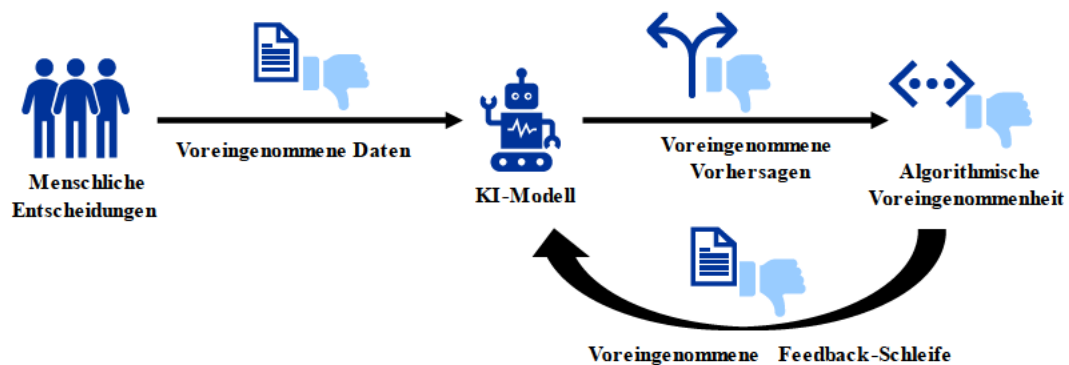


Abbildung 12: Feedback-Schleife (eigene Darstellung) (Marques, 2021)

Wie im Unterkapitel „Entstehung eines KI-Modells“ bereits illustriert, erzeugen menschliche Handlungen voreingenommene Daten. Eine Teilmenge derjenigen Daten wird in Form von Trainingsdaten dem KI-Modell initial zum Erlernen seines Verhaltens zugespielt und würde somit das Modell für ein ganzes Leben prägen. Dies hat den Grund, dass das Verhalten, welches das Modell anfänglich selbstständig erlernte, seine ersten Entscheidungen bestimmt und diejenigen Entscheidungen dann in Form von neuen Daten in den Trainings- und Testdatensatz einfließen und somit das Verhalten des Modells verfestigen (Machill, 2020; Marques, 2021).

Dies könnte konkret so aussehen, dass das Modell auf das Amazon-Beispiel bezogen, initial erlernt, Frauen zu benachteiligen und Männer zu bevorzugen. Folglich würden bei gleicher fachlicher Qualifikation der zwei unterschiedlichen Geschlechter trotzdem nur Männer vom Modell als Top-Kandidaten ausgewählt und zurückgeliefert werden.

Letzten Endes könnte die Empfehlung des Modells darin münden, dass auch nur Männer von Amazons Recruitern eingestellt werden. Die Daten, also die möglichen Kandidaten und wer von ihnen final eingestellt worden ist, würden als neue Daten in den Datensatz einfließen, mit welchem das Modell trainiert und validiert wird. Anschließend würde das KI-Modell in der nächsten Lern-Iteration auf Grundlage dieses erweiterten voreingenommenen Datensatzes eigenständig sein bisheriges voreingenommenes Verhalten verstärken, da die neuen Daten auf dem initial erlernten Verhalten beruhen und dieses somit bestätigen sowie vertiefen würden (Machill, 2020; Marques, 2021).

Die erzeugten Ergebnisse des Systems würden also als positives Feedback für das Verhalten des Modells fungieren und im Effekt dem KI-Modell versichern, dass er sich richtig verhält, was im Ergebnis darin mündet, dass sich sein Verhalten verfestigt und intensiviert (Machill, 2020; Marques, 2021).

Diese Feedback-Schleife illustriert, wie sich algorithmische Voreingenommenheit eigenständig verstärken kann und im Resultat diejenigen Betroffenen des Systems zunehmend diskriminiert (Machill, 2020; Marques, 2021).

Eine Alternative zu KI-Modellen, die sich mithilfe von Feedback-Schleifen „weiterentwickeln“, wären *KI-Modelle, die sich nie verändern*, sondern immer auf dem initialen Verhaltensstand bleiben (Machill, 2020).

Auch wenn Feedback-Schleifen, wie hier vorgestellt, bestehende algorithmische Voreingenommenheiten vertiefen könnten, geben sie einem Modell ebenfalls die Möglichkeit, sich zu verbessern und somit unter anderem Voreingenommenheiten zu mildern. Diese Verbesserung des Modells kann nur stattfinden, wenn keine neuen voreingenommenen Daten in den Trainingsdatensatz einfließen (Machill, 2020; Marques, 2021).

Nichtsdestotrotz hätten sich nie verändernde KI-Modelle keine Möglichkeit der Verbesserung. Diese würden unabhängig von Veränderungen der Realität agieren und somit ihre eigene schaffen, in welcher sie ihr initial erlerntes Verhalten verfolgen. Dies hätte zur Folge, dass fehlerhafte Arbeitsweisen oder initial erlernte algorithmische Voreingenommenheiten für immer im Verhalten des KI-Modells bestehen bleiben d. h. bestimmte Personengruppen werden für immer diskriminiert und fehlerhafte Arbeitsweisen bleiben ewig im System (Machill, 2020).

Mit der vorhandenen oder nicht vorhandenen Feedback-Schleife wurden zwei Möglichkeiten vorgestellt, wie AES ihr voreingenommenes oder unvoreingenommenes Verhalten verstärken könnten.

### **9.3.3 Korrelation und Kausalität**

Eine weitere mögliche Ursache für algorithmische Voreingenommenheit ist mit den Begriffen *Korrelation und Kausalität* vorhanden (Machill, 2020).

Wie an einigen Stellen dieser Arbeit bereits erwähnt wurde, nutzen KI-Algorithmen Korrelationen, um Prognosen (besser) treffen zu können. Mit Korrelationen werden bestimmte Phänomene bezeichnet, die eine wechselseitige Beziehung aufweisen. Beispielsweise korreliert die Übergewichtigkeit eines Menschen mit seiner Lebenserwartung, es existieren statistische Hinweise darauf, dass Übergewichtige eventuell eine kürzere Lebensdauer haben als Normalgewichtige. Nichtsdestotrotz existiert hier keine Kausalität. Denn es ist nicht zwangsläufig so, dass übergewichtige Menschen frühzeitiger sterben. Nichtsdestotrotz könnte ein KI-Algorithmus anhand des Gewichtes einer Person eine Vorhersage darüber treffen, wie lange diese wahrscheinlich noch leben wird. Also werden Korrelationen demzufolge aufgrund ihrer potenziellen Prognosefähigkeit von KI-Modellen genutzt (*Korrelation als Kausalität*, o. D.; Machill, 2020).

Der Unterschied zwischen Korrelation und Kausalität ist also wichtig. Bei einer Korrelation steht das eine „[...] in Beziehung zum anderen, bedingt es aber nicht zwingend“ (*Korrelation als Kausalität*, o. D.). Während bei einer Kausalität eine Ursache-Wirkung-Beziehung vorherrscht, dies bedeutet „Das eine [*sic*] verursacht das andere [*sic*]“ (*Korrelation als Kausalität*, o. D.) (*Korrelation als Kausalität*, o. D.; Machill, 2020).

KI-Modelle können jedoch nicht zwischen diesen beiden Begriffen unterscheiden. Dementsprechend könnte ein KI-Modell von einer Kausalität zwischen zwei Variablen ausgehen und innerhalb des Algorithmus dementsprechende Regeln formulieren, wobei nur maximal eine Korrelation vorliegt, sodass diese Regeln nicht der Realität entsprechen. Hierzu ein bekanntes Beispiel, welches die genaue Analyse zwischen Merkmalen hervorhebt, bevor man darüber entscheidet, ob diese miteinander korrelieren oder sogar einen kausalen Zusammenhang haben (*Korrelation als Kausalität*, o. D.; Machill, 2020).

Im Jahr 1940 wurde aufgrund der Gefahr des Poliovirus, welches Kinder paralisieren könnte, davon abgeraten, Eiscreme von mobilen Eiscreme-Verkäufern zu konsumieren, da sich das Virus per Schmierinfektion weiterverbreiten kann. Es wurde eine Korrelation zwischen der

steigenden Anzahl an Polioerkrankten und der steigenden Anzahl an Eiscreme-Verkäufen vermutet (*Kinderlähmung (Poliomyelitis)*, 2020; Machill, 2020).

Ein KI-Modell könnte auf Grundlage der beiden steigenden Merkmale ebenfalls solch eine Korrelation oder sogar Kausalität feststellen und in seinen Handlungsregeln verfestigen. Dies könnte so aussehen, dass wenn das KI-System dafür ausgelegt wäre, Krankheiten bei Kindern zu prognostizieren, dann würde es bei einem Kind, von welchem bekannt ist, dass es viel Eiscreme konsumiert, möglicherweise die Vorhersage treffen, dass dieses Kind auch von Polio erkrankt sein könnte. Allerdings existiert kein medizinischer Zusammenhang zwischen diesen beiden Merkmalen, die Korrelation, die eigentlich vorliegt, wäre die Korrelation der beiden Merkmale mit der Jahreszeit Sommer (Machill, 2020).

Dieses Beispiel illustriert, wie ein KI-Modell aufgrund von Identifizierung nicht vorhandener Zusammenhänge fehlerhafte Annahmen treffen könnte und welche Auswirkungen solche verzerrten Annahmen haben könnten (Machill, 2020).

Dieses Unterkapitel stellt mit Korrelation und Kausalität, zwei für KI-Systeme relevante Begrifflichkeiten vor. Weiterführend illustriert das hier vorgestellte Beispiel, dass die algorithmische Voreingenommenheit sich nicht unbedingt auf das Verletzen von geschützten Merkmalen beziehen muss, sondern man auch von algorithmischer Voreingenommenheit sprechen kann, wenn der erlernte Realitätsausschnitt eines KI-Algorithmus verzerrt ist. Wenn also wie hier das KI-System aufgrund der Ermittlung von bedeutungslosen Zusammenhängen ein fehlerhaftes bzw. verzerrtes Bild der Realität verfolgt, welches wie dargestellt in beispielsweise gesundheitlichen Prognosen münden kann, die fehlerhaft sind (Machill, 2020).

### **9.3.4 Proxy-Variablen**

Eine weitere, ebenfalls auf Korrelation basierende Ursache für algorithmische Verzerrung, liegt mit *Proxy- bzw. Platzhalter-Variablen* vor. Sollten einem KI-Modell bestimmte Variablen nicht vorliegen, welche es für seine Entscheidungsfindung benötigt, so greift es auf vorhandene Daten zurück und versucht die fehlende Information abzuleiten. Diejenige Information, aus welcher die fehlende abgeleitet werden kann, wird Proxy- bzw. Platzhalter-Variable genannt.

Solche Platzhalter-Variablen werden vom KI-Modell insbesondere dann verwendet, wenn geschützte Merkmale exkludiert werden, um möglichen Voreingenommenheiten des KI-Systems vorzubeugen. Im Grunde also um zu verhindern, dass das KI-Modell diese geschützten Merkmale auf diskriminierende Art und Weise, wie in den vielen Fallbeispielen illustriert, für seine Entscheidungsfindung berücksichtigt. Dieses Vorhaben könnte jedoch zur Folge haben, dass das KI-System aufgrund der Proxys trotzdem diskriminierend sowie voreingenommen ist und darüber hinaus sind die durch Platzhalter entstehenden Voreingenommenheiten schwieriger zu ermitteln und bekämpfen, als wenn man direkt geschützte Variablen einbezogen hätte. Dies lässt sich anhand des COMPAS-Fallbeispiels näherbringen (De Montis, 2019; Marques, 2021; Mohammad, 2021).

Wie bereits im Unterkapitel von COMPAS erörtert, diskriminiert das KI-System dunkelhäutige Menschen, obwohl es keine Informationen über die geschützten Merkmale wie Hautfarbe oder Ethnie enthält. Das KI-Modell greift nämlich auf andere bestehende Variablen zurück, die mit der Hautfarbe und Ethnie einer Person statistisch gesehen in Zusammenhang stehen. Eine in Datensätzen häufig anzutreffende Variable wäre die Postleitzahl einer Person. Anhand dieser wäre ein Rückschluss auf die Hautfarbe und Ethnie eines Menschen zurückführbar, da Postleitzahlen historisch gesehen stark mit der ethnischen Zusammensetzung von Wohngebieten korrelieren. Mithilfe dieser Proxy-Variable könnte das KI-Modell also trotzdem rassistische Diskriminierung ausüben. Im Resultat könnte ein neutral und harmlos wirkendes Merkmal wie die Postleitzahl für algorithmische Ungleichbehandlung von Minderheiten sorgen (Kolleck & Orwat, 2020; Marques, 2021; Mohammad, 2021; Rätz, 2022).

Nun könnte man überlegen, ob man diese Proxy-Variablen ebenfalls aus dem Datensatz des KI-Modells entfernt, sodass auch keine indirekten Indikatoren auf jegliche geschützten Variablen vorliegen und das KI-System somit fortan ohne Voreingenommenheit Entscheidungen trifft. Nichtsdestotrotz würde man somit immer mehr Informationen aus dem Modell entfernen und würde somit eventuell Variablen entfernen, die die korrekte Funktionsweise des Systems beeinflussen und somit dessen Performance sowie Genauigkeit beeinträchtigen. Weiterführend wird angenommen, dass bei Entfernung einer Proxy-Variablen andere Platzhalter für diesen Platzhalter einspringen könnten. Im vorgestellten COMPAS-Beispiel würde dies bedeuten, dass bei Entfernung der Ethnie, Hautfarbe und Postleitzahl eine andere neutral erscheinende



Variable als Platzhalter für die Postleitzahl und somit ebenfalls als Platzhalter für die geschützten Variablen fungiert (Machill, 2020; Mohammad, 2021).

Es wird deutlich, dass das Entfernen von geschützten Merkmalen, algorithmische Voreingenommenheit und Diskriminierung nicht verhindert oder mindert. Denn KI-Modelle sind in der Lage, Zusammenhänge in den Daten zu erkennen und somit auf Proxy-Variablen zurückzugreifen, die als Platzhalter für diese geschützten Variablen dienen, und somit eventuell die algorithmische Voreingenommenheit sowie Diskriminierung des KI-Modells verschlimmern sowie aufwändiger zu ermitteln machen (De Montis, 2019; Marques, 2021; Mohammad, 2021).

### **9.3.5 Über- und Unterrepräsentierung**

Eine weitere mögliche Ursache von algorithmischer Verzerrung liegt vor, wenn ein Ungleichgewicht der Daten im Datensatz vorherrscht. Ein KI-Modell arbeitet umso korrekter und genauer, umso mehr adäquate Daten vorliegen. Sollte es nun so sein, dass äußerst viele Daten der Mehrheit und äußerst wenige der Minderheit vorliegen, so kann es passieren, dass das KI-Modell Entscheidungen, welche die Mehrheit betreffen, sehr genau, korrekt und performant ermitteln kann, während Entscheidungen, welche die Minderheit betreffen, fehleranfälliger sind (Machill, 2020; Mohammad, 2021).

Nun ist es so, dass Chen et al. (2018) in ihrer Recherche zu dem Entschluss gekommen sind, dass KI-Modelle für Minderheiten eine gleich akkurate Vorhersagegenauigkeit wie für die Mehrheit benötigen, um algorithmische Diskriminierung sowie Voreingenommenheit zu vermeiden (Machill, 2020).

Diese *Unterrepräsentierung* kam bereits beim Amazon-Fallbeispiel zum Vorschein, lässt sich jedoch ebenfalls anhand eines weiteren Beispiels verdeutlichen (Machill, 2020; Mohammad, 2021).

Dieses Beispiel befasst sich mit KI-Gesichtserkennungssoftware. Diese KI-Systeme werden dazu verwendet, Gesichter zu erkennen sowie zu klassifizieren. Nichtsdestotrotz fehlklassifizieren sie dunkelhäutige Gesichter häufiger als hellhäutige Gesichter. Die zugrundeliegende Verteilung an hellhäutigen und dunkelhäutigen Gesichtern in den Trainingsdaten könnte

hierfür verantwortlich sein, da schätzungsweise rund 84% der Gesichter in den aktuell verwendeten Trainingsdatensätzen für solche Software hellhäutig sind. Dementsprechend wären dunkelhäutige Personen in diesen Datensätzen unterrepräsentiert, sodass folglich zunehmend diese Minderheit fehlerklassifiziert wird, was in algorithmischer Diskriminierung gegenüber dieser mündet (Machill, 2020).

In der Strafverfolgungsbehörde der USA wird ebenfalls KI-Gesichtserkennungssoftware eingesetzt. Berücksichtigt man nun die Tatsache, dass diese KI-Systeme aufgrund der in den USA vorherrschenden Diskriminierung gegenüber dunkelhäutigen Personen zum Großteil für dunkelhäutige Menschen eingesetzt werden, so wird klar, dass diejenigen, die am häufigsten von diesen KI-Modellen klassifiziert werden, diejenigen sind, die die höchste Fehlerklassifizierung zu erwarten haben. Im Ergebnis würde dies in Verhaftungen und eventuell sogar Verurteilungen von den falschen bzw. unschuldigen Personen münden (Machill, 2020; Mohammad, 2021).

Mit diesem Beispiel wurde illustriert, wie die Unterrepräsentierung gewisser Daten wie Daten der Minderheiten in algorithmische Verzerrung münden kann und wie diskriminierend die Auswirkungen für die betroffenen Personen sein können.

Um die Auswirkungen von *Überrepräsentierung* zu veranschaulichen, wird das im Unterkapitel „Rechtswesen“ vorgestellte KI-System für prädiktive Polizeiarbeit herangezogen. Dort wurde bereits erläutert, wie historische Daten und Feedback-Schleifen dieses KI-Modell diskriminierend gegenüber dunkelhäutigen sowie hispanischen Personen macht. Im Ergebnis könnten die Daten dieser Minderheiten im Datensatz des KI-Modells überrepräsentiert sein. Im Kontext dieses KI-Systems würde dies dafür sorgen, dass sich das KI-Modell auf diese Minderheiten fokussiert und zunehmend prädiktive Polizeiarbeit über diese ergehen lässt (Mohammad, 2021).

Abschließend ist es wichtig anzusprechen, dass diese Über- und Unterrepräsentierung der Daten, ein Indikator für fehlende Diversität und Ausgewogenheit des Datensatzes ist (Gasser, 2019).

Hiermit konnte die Über- und Unterrepräsentierung sowie die daraus resultierenden möglichen algorithmischen Voreingenommenheiten verdeutlicht werden.

## **9.4 KI-Entwickler und Situation der KI-Branche**

Im Folgenden Unterkapitel wird erst einmal vermittelt, wer die Entwickler dieser KI-Modelle sind, welche Aufgaben sie haben und wie ihre Voreingenommenheiten in die Entwicklung und folglich auch in das Verhalten der Modelle einfließt. Daraufhin wird die aktuelle Lage der KI-Branche thematisiert, wobei insbesondere auf die mangelnde Diversität der in dieser Branche tätigen Entwickler eingegangen wird.

### **9.4.1 Datenwissenschaftler (Data Scientist)**

Die Entwickler von KI-Modellen werden auch *Datenwissenschaftler oder Data Scientists* genannt. „Ein Data Scientist oder Datenwissenschaftler ist ein Analytiker, der für das Sammeln, Analysieren und Interpretieren von Daten verantwortlich ist [...]“ (Fitzgibbons, Laskowski & Botelho, 2022). Diese gesammelten und aufbereiteten Daten werden unter anderem in Form von Trainingsdatensätzen dem KI-Modell zum Erlernen seines Verhaltens zur Verfügung gestellt (Bär, 2022; Fitzgibbons et al., 2022).

Menschen sind voreingenommen und KI-Modelle werden letzten Endes von Datenwissenschaftlern, d. h. Menschen konzipiert und entwickelt. Dementsprechend können deren Voreingenommenheiten in das Verhalten des KI-Modells einfließen (Bär, 2022; Marques, 2021).

Bei der Gestaltung dieser Modelle muss unter anderem darüber entschieden werden, welche Daten man als Trainings- und Testdaten heranzieht, welche Merkmale man im jeweiligen KI-Modell berücksichtigt, ebenso deren Gewichtung. Wie diese Merkmale gemessen werden sollen sowie wie die Ziel- und Klassenvariablen definiert werden sollen. Bei diesen Überlegungen können Voreingenommenheiten des Entwicklers oder bestimmte Designentscheidungen in verzerrten KI-Modellen münden (Bär, 2022; Marques, 2021; Mohammad, 2021).

### **Bestätigungsvoreingenommenheit bei der Wahl des Trainingsdatensatzes**

Der Datenwissenschaftler entscheidet darüber, aus welchen Daten sich der Trainings- sowie Testdatensatz bildet. Hierauf kann die in Kapitel „Ursachen kognitiver Verzerrung“ bereits vorgestellte Bestätigungsvoreingenommenheit des Entwicklers Einfluss nehmen (Bär, 2022).

Bei dieser kognitiven Verzerrung ist es in diesem Kontext so, dass das bereits bestehende Meinungsbild eines Entwicklers sich darauf auswirkt, welche Daten er für den Trainings- und Testdatensatz auswählt. Angenommen, der Entwickler wäre unterbewusst voreingenommen gegenüber Frauen. Dann könnte er Frauen, eventuell ungewollt, im Trainingsdatensatz unterrepräsentieren und würde somit dafür sorgen, dass das KI-Modell voreingenommen und eventuell sogar diskriminierend gegenüber Frauen ist. Mögliche konkrete Auswirkungen eines solchen Falls wurden mit dem Amazon-Fallbeispiel bereits illustriert (Bär, 2022).

### **Wahl und Gewichtung der Merkmale**

Damit ein KI-Modell zu Ergebnissen kommen kann, benötigt es Merkmale, die es in die Entscheidungsfindung einbeziehen kann, um somit akkurate und korrekte Ergebnisse liefern zu können. Nichtsdestotrotz ist hier darauf zu achten, welche Merkmale man als Entwickler dem KI-Modell übergibt. Eine ausgewählte Menge an Merkmalen kann den Rechen- sowie Speicheraufwand des KI-Systems reduzieren und somit für eine schnellere Verarbeitung und Entscheidungsfindung des Modells sorgen (*Feature Selection*, 2022; Mohammad, 2021).

Ziel des Entwicklerteams ist es, eigenständig oder im Austausch mit den Auftraggebern diejenigen Variablen zu identifizieren, die zur Verbesserung der Vorhersagegenauigkeit des Modells beitragen und somit die überflüssigen Merkmale aus dem Modell zu entfernen. Auf diese Weise würde die Komplexität des KI-Modells sinken und die Verständlichkeit der Entscheidungsfindung sowie die Nachvollziehbarkeit der finalen Entscheidung steigen (*Feature Selection*, 2022; Mohammad, 2021).

Nichtsdestotrotz kann die Auswahl der Merkmale darin münden, dass bestimmte Personen, unter anderem Minderheiten, diskriminiert werden (Mohammad, 2021).

Hierfür kann man sich beispielhaft ein KI-System vorstellen, welches einem Unternehmen aus den USA beim Einstellungsprozess helfen soll. Angenommen, die Recruiter legen großen Wert darauf, wo der Bewerbende seinen Abschluss gemacht hat bzw. welchen Ruf die jeweilige Universität aufweist. Dann würden die Entwickler im Austausch mit den Recruitern das Merkmal „Ruf der Universität“ im KI-Modell aufnehmen oder beibehalten und diesem eine hohe Gewichtung zuschreiben. Die Recruiter gehen nämlich davon aus, dass Absolventen einer Universität mit gutem Ruf äußerst talentiert sind und sich somit sehr gut als Mitarbeiter im Unternehmen eignen könnten (Mohammad, 2021).

Nun ist die Situation, insbesondere in den USA, jedoch so, dass das Studieren an einer guten Universität mit hohen Studiengebühren verbunden ist (Mohammad, 2021; *Was kostet ein Studium im Ausland?*, o. D.).

Dies hätte zur Folge, dass Studenten aus einem armen oder mittelständigen Haushalt nur mit einem Stipendium an solchen Universitäten studieren können. Sodass man erwarten kann, dass die Absolventen dieser Universitäten eher aus einem reichen Haushalt stammen (Economic diversity and student outcomes at Harvard University, 2017; Mohammad, 2021).

In den USA wären dies statisch gesehen eher hellhäutige Familien, während diejenigen, die nicht genug Geld für solche Bildungsstätten verfügen, eher einen Migrationshintergrund aufweisen und eher als Minderheit identifiziert werden (Meschede, 2016; Mohammad, 2021).

Dementsprechend würde die Berücksichtigung und Gewichtung des Merkmals „Ruf der Universität“ nicht darin münden, dass unbedingt die besten Kandidaten für freie Stellen am Unternehmen vom AES ausgewählt werden, sondern eher darin, dass diejenigen außer Acht gelassen werden, die fachlich eventuell gleich oder sogar besser qualifiziert wären als Absolventen einer guten Universität. Oftmals wären die außer Acht gelassenen Kandidaten Menschen einer Minderheit (Mohammad, 2021).

Mit diesem Beispiel wird illustriert, wie wichtig sowie fragil die Auswahl und Gewichtung von Merkmalen für KI-Systeme ist. Weiterführend wird die Relevanz hervorgehoben, genau zu untersuchen, welche Auswirkungen und welchen Beitrag zur Verbesserung der Genauigkeit ein Merkmal letzten Endes mit sich bringt und wie neutral es ist. Es wurde nämlich deutlich, dass die Berücksichtigung und hohe Gewichtung von falschen Merkmalen das Verhalten eines

AES so massiv beeinflussen kann, dass die von diesem KI-Modell erschaffenen Entscheidungen stark von diesem Merkmal abhängig sind und eventuell diskriminierend gegenüber bestimmten Personen sein können (Mohammad, 2021).

### **Messverzerrung**

Bei manchen Merkmalen muss überlegt werden, wie man diese am besten ermittelt bzw. misst. Das Merkmal „Ruf der Universität“ aus dem soeben thematisierten Beispiel kann mit wenig Aufwand ermittelt werden. Nichtsdestotrotz existieren auch Merkmale, bei welchen die Ermittlung nicht so mühelos vollzogen werden kann. Ein Beispiel für solch ein Merkmal lässt sich anhand des COMPAS-Beispiels anführen. Bei diesem liegt eine Variable „Risikofreudigkeit“ vor, welche wiedergeben soll, wie risikoreich es ist, dass die jeweilige Person rückfällig wird und somit eine weitere Straftat durchführen wird. Hier haben die Entwickler sich dazu entschieden, frühere Festnahmen und Festnahmen in der Familie als Messgrößen für diese Variable zu nehmen. Somit würde das KI-Modell bei der Ermittlung dieses Merkmals eine Messverzerrung aufweisen, welche in algorithmische Voreingenommenheit gegenüber Afro- und Hispanoamerikanern mündet (Marques, 2021; Mohammad, 2021).

Um dies nachvollziehen zu können, müsste man sich an die Tatsache erinnern, dass diese Minderheiten in den USA systematisch diskriminiert werden und beide zusammen weit mehr als die Mehrheit an Inhaftierten in den USA repräsentieren (*Anzahl der Inhaftierten in den USA nach Ethnie im Jahr 2021*, 2023).

Hierzu wurde ebenfalls im Unterkapitel „Rechtswesen“ von einem AES berichtet, welches gezielt vermehrt prädiktive Polizeiarbeit auf Wohnorte dieser Minderheiten richtet. Dementsprechend kommen innerhalb dieser Wohngegenden auch mehr unberechtigte und falsche Festnahmen sowie Verurteilungen vor (*Schwarze häufiger unschuldig verurteilt als Weiße*, 2022).

Auf Grundlage dieses Wissens und der Tatsache, dass jeder dritte Afroamerikaner und zusätzlich jeder sechste Hispanoamerikaner eines Tages inhaftiert wird, liegt es nahe, davon auszugehen, dass in einer afro- oder hispanoamerikanischen Familie mindestens ein Mitglied dieser Familie höchstwahrscheinlich eine kriminelle Vergangenheit aufweist oder momentan sogar im Gefängnis seine Haftstrafe absitzt.

Dementsprechend würde das Merkmal „Risikofreudigkeit“ aufgrund der Art und Weise seiner Ermittlung bei Personen dieser Minderheiten dahingehend tendieren, eher ein hohes Rückfallrisiko für diese vorherzusagen. Also könnten ein dunkelhäutiger Afroamerikaner und ein hellhäutiger Amerikaner die gleiche Straftat und Strafhistorie aufweisen, jedoch würde das KI-Modell dem Afroamerikaner lediglich aufgrund seiner Familie bzw. seines Wohnortes mit erhöhter Wahrscheinlichkeit ein höheres Rückfälligkeitsrisiko zuschreiben als dem hellhäutigen Amerikaner.

Hierbei wurde noch nicht einmal mitberücksichtigt, dass Afroamerikaner statisch gesehen eher eine kriminelle Vergangenheit aufweisen als hellhäutige Amerikaner und somit wahrscheinlich auch mehr historische Festnahmen aufweisen, auch wenn diese unberechtigter Natur sein können. Also würde das AES eventuell zusätzlich zu den Festnahmen in der Familie diese historischen Festnahmen ebenfalls negativ in die Ermittlung des Merkmals „Risikofreudigkeit“ berücksichtigen (Machill, 2020; *Schwarze häufiger unschuldig verurteilt als Weiße*, 2022).

Mit diesem auf COMPAS aufgebauten Beispiel konnte illustriert werden, was man unter Messverzerrung verstehen kann und wie diese in algorithmische Voreingenommenheit münden kann.

### **Definition der Eigenschaften der Zielvariablen**

Wenn ein Entwicklerteam den Auftrag bekommt, ein AES für den Auftraggeber zu erstellen, so muss der Auftraggeber diesem Team mitteilen, was das Ziel des finalen KI-Systems sein soll. Dieses Ziel könnte beispielsweise sein, dass das KI-Modell letzten Endes in der Lage ist, „gute Mitarbeiter“ zu identifizieren. Die Definition der Zielvariable wäre also „guter Mitarbeiter“. Nun ist es jedoch so, dass die Entwickler nicht eindeutig darüber aussagen können, wer solch ein Mitarbeiter ist. „Gut“ ist nämlich keine objektive Eigenschaft, sondern ist eine subjektive Wahrnehmung (Mohammad, 2021).

Beispielsweise wären top Verkaufszahlen eines Mitarbeiters zwar objektiv messbare Kennzahlen, diese allein reichen jedoch nicht aus, um einen Mitarbeiter als einen „guten Mitarbeiter“ zu klassifizieren. Dementsprechend muss das Entwicklerteam, eventuell gemeinsam mit dem Auftraggeber eine einheitliche Definition solch Mitarbeiter entwickeln. Diese Definition wäre

dann wie jegliche Definition einer Ausprägung einer möglichen Zielvariable in den Klassenbezeichnungen in Form einer Kategorie niedergelegt. Im Grunde gäbe es in den Klassenbezeichnungen somit eine Kategorie „gut“. Damit ein Mitarbeiter in diese Kategorie fällt, müsste er die mit dieser Kategorie verbundenen Bedingungen erfüllen (Mohammad, 2021).

Zusätzlich zu objektiv messbaren Kennzahlen wie die Verkaufszahlen eines Mitarbeiters wären subjektive Eigenschaften wie beispielsweise die Teamfähigkeit oder Zuverlässigkeit in Form von Pünktlichkeit heranzuziehen. Man könnte darüber debattieren, ob die Pünktlichkeit eines Mitarbeiters nicht auch eine objektive Kennzahl ist, jedoch ist es so, dass man hier ebenfalls berücksichtigen könnte, welchen Aufwand und wie viel Einfluss eine Person darauf hat, ob sie zu spät zur Arbeit erscheint oder nicht (Mohammad, 2021).

Hierzu ein Beispiel. Angenommen es existieren zwei Mitarbeiter für ein Unternehmen, dessen Büro in der Stadt-Mitte lokalisiert ist und es soll mithilfe eines AES ermittelt werden, welcher dieser beiden eher gefeuert werden sollte (Mohammad, 2021).

Einer dieser Mitarbeiter wohnt fünf Minuten Fußweg von der Arbeit entfernt und ist insgesamt fünf Mal zu spät im Büro erschienen. Auf der anderen Seite hat man einen Mitarbeiter, der eine Stunde und 15Min. Fahrt benötigt, um bei der Arbeit zu erscheinen. Dieser muss auf Öffentliche Verkehrsmittel zugreifen, da er eventuell aus finanziellen Gründen außerhalb der Stadt wohnt. Dieser Mitarbeiter kam sieben Mal zu spät zur Arbeit. Objektiv gesehen, wäre der erste Mitarbeiter der Zuverlässigere der beiden, da er weniger häufig zu spät zur Arbeit erschien als der zweite Mitarbeiter. Berücksichtigt man jedoch, dass bei einem so langen Weg zur Arbeit, insbesondere mit öffentlichen Verkehrsmitteln, vieles nicht nach Plan verlaufen kann, dann wird man als Arbeitgeber eventuell doch den zweiten Mitarbeiter als zuverlässiger einstufen, da dieser - seinen Umständen entsprechend - verlässlicher pünktlich zur Arbeit erscheint (Mohammad, 2021).

Angenommen die Entwickler entscheiden sich zusammen mit dem Auftraggeber dafür, bei der Zuverlässigkeit bzw. Pünktlichkeit, den objektiven Ansatz zu verfolgen. Mit der Wahl dieses Ansatzes würde das AES wahrscheinlich eher Mitarbeiter bevorzugen, die innerhalb der Stadt wohnen, da diese mehr Kontrolle und Planungskraft über ihren Arbeitsweg verfügen und somit wahrscheinlicher pünktlich zur Arbeit erscheinen werden als Mitarbeiter, die außerhalb der Stadt wohnen. Berücksichtigt man die Tatsache, dass Grundstücke innerhalb der Stadt generell



teurer sind als außerhalb und dass Menschen, die außerhalb der Stadt wohnen, dies somit eventuell aus Kostengründen tun, so wird deutlich, dass das AES durch eine Bevorzugung von Menschen, die innerhalb der Stadt wohnen, eine algorithmische Voreingenommenheit gegenüber finanziell ärmeren oder sparsameren Mitarbeitern aufweisen könnte. Je nach Gewichtung dieser Eigenschaft könnte dies im Extremfall dafür sorgen, dass fast alle Mitarbeiter, die innerhalb der Stadt wohnen, als „gute Mitarbeiter“ klassifiziert werden, während nahezu jeder Mitarbeiter, welcher außerhalb der Stadt wohnt, nicht als solch ein Mitarbeiter klassifiziert wird. Konkret könnte dies bedeuten, dass der zweite Mitarbeiter, welcher außerhalb der Stadt wohnt und subjektiv gesehen eventuell verlässlicher pünktlich zur Arbeit erscheint als der erste Mitarbeiter, nicht als „guter Mitarbeiter“ klassifiziert wird und somit im Extremfall aufgrund des AES gefeuert wird (Mohammad, 2021).

Mit diesem Beispiel wird illustriert, dass Entwickler sowie Auftraggeber stark darauf zu achten haben, welche Definitionen sie in den Klassenbezeichnungen niederlegen, da dies direkte oder indirekte Auswirkungen auf das Verhalten des AES hat und weiterführend in algorithmische Voreingenommenheit bzw. diskriminierendem Verhalten münden kann (Mohammad, 2021).

### **Fehlende Diversität**

Im Folgenden wird auf die *fehlende Diversität der Arbeitenden innerhalb der Technologiebranche*, insbesondere auf die ausbleibende Vielfalt der Entwickler innerhalb der KI-Branche näher eingegangen und es wird hervorgehoben, wie diese mangelnde Diversität in algorithmische Voreingenommenheit münden kann (De Montis, 2019; Machill, 2020).

Die Technologiebranche und insbesondere der Teilbereich der künstlichen Intelligenz wird von einer bestimmten Personengruppe dominiert. Diese homogene Gruppe besteht aus hellhäutigen Männern (De Montis, 2019; Machill, 2020).

Dies spiegelt sich in den KI-Forschungsteams einiger der größten Technologieunternehmen wie Facebook oder Google wider. Bei diesen beiden Tech-Giganten würde das KI-Forschungsteam aus mindestens 85% Männern bestehen, wobei nur maximal 4% der eingestellten Personen dunkelhäutig ist (*Aufstieg der Tech-Giganten*, 2021; Machill, 2020).

Wie bereits illustriert ist jeder Mensch zu einem gewissen Grad voreingenommen. Nun ist es jedoch so, dass diese Voreingenommenheit je nach Individuum unterschiedliche Ausprägungen annehmen kann. Nichtsdestotrotz könnten homogene Gruppen wie hier hellhäutige Männer ähnliche Voreingenommenheiten aufweisen und diese beabsichtigt oder unbeabsichtigt in das AES einfließen lassen. Aufgrund der Tatsache, dass diese eventuell in vielen Punkten ein ähnliches Weltbild verfolgen, könnte es solch einer homogenen Gruppe schwerfallen, eigene Vorurteile zu erkennen und darauf zu achten, dass diese keinen Einfluss auf die Entwicklung des AES nehmen. Im Grunde könnte einem zu homogenen Entwicklerteam also ein Bewusstsein für mögliche Voreingenommenheiten fehlen (De Montis, 2019; Machill, 2020; Marques, 2021).

Wenn die Entwickler des KI-Systems ihre Voreingenommenheiten nicht erkennen und bekämpfen oder mindern, so würden voreingenommene historische Daten ebenfalls nicht erkannt und gemindert werden. Beispielsweise könnte es sein, dass ein Team welches ausschließlich aus männlichen Entwicklern besteht, während der Entwicklung eines AES keine voreingenommenen historischen Daten gegenüber Frauen erkennt (De Montis, 2019; Machill, 2020; Marques, 2021).

Dies möglicherweise aus dem Grund, dass die Entwickler eventuell ebenfalls unterbewusst Frauen gegenüber voreingenommen sind. Solch eine aufkommende algorithmische Voreingenommenheit innerhalb eines AES konnte mit dem Amazon-Fallbeispiel bereits verdeutlicht werden. Hier ist dem KI-Entwicklerteam nicht aufgefallen, dass die verwendeten Trainingsdaten Frauen gegenüber historisch voreingenommen waren (De Montis, 2019; Machill, 2020; Marques, 2021).

Hier hätte das Einbeziehen von Frauen in das Entwicklerteam vorbeugend wirken können und eventuell hätte man somit diskriminierende Entscheidungen des AES gegenüber Frauen vermeiden können. Auf diversere Teams wird in den Lösungen vertiefter eingegangen (De Montis, 2019; Machill, 2020; Marques, 2021).

Dieses Unterkapitel illustriert jedoch, wie die fehlende Diversität eines Entwicklerteams in algorithmische Voreingenommenheit münden kann. Weiterführend wird hervorgehoben, dass die Mehrheit der KI-Entwickler zu einer homogenen Personengruppe gehört. Dies bedeutet, dass diejenigen mächtigen KI-Systeme und AES die Urteile und Vorhersagen zu jeglichen

Personengruppen (Frauen, dunkelhäutige, homosexuelle usw.) treffen, letzten Endes von einer privilegierten Personengruppe entwickelt werden. Dies spiegelt sich, wie mit dem Amazon-Beispiel aufgezeigt, im Verhalten der KI-Modelle wider. Hierbei müssen die Entwickler noch nicht einmal explizit voreingenommen sein, sondern es reicht schon, wenn diese unterbewusst voreingenommen sind und ihnen das Bewusstsein für diese Vorurteile fehlt (Bär, 2022; De Montis, 2019; Machill, 2020; Marques, 2021).

### **Ego-Depletion**

Eine weitere Ursache algorithmischer Voreingenommenheit hängt wieder eng mit der kognitiven Arbeitsweise des Menschen zusammen. Wie im Kapitel „Ursachen kognitiver Verzerrung“ bereits erörtert, würde unser Gehirn Vorurteile verwenden, um Aufgaben energieeffizienter bewältigen zu können. Dementsprechend liegt es nahe, dass insbesondere dann, wenn der Entwickler viele feingranulare Entscheidungen getätigt hat oder bereits seit einigen Stunden an der gleichen Thematik arbeitet, dass sein Gehirn folglich nicht mehr über genug Energie verfügt, um jede Entscheidung bis ins Detail zu durchdenken und objektiv zu bleiben (Bär, 2022).

Vielmehr würde das Gehirn, um die Aufgaben bewältigen zu können, zunehmend auf Voreingenommenheiten zurückgreifen. Besonders Vorurteile, die in diskriminierenden Ergebnissen münden könnten, würden an dieser Stelle in die Entwicklung des KI-Systems einfließen. Dementsprechend befindet sich der Entwickler während der *Ego-Depletion* in einer Phase der kognitiven Erschöpfung. In dieser würde sein Gehirn vermehrt Voreingenommenheiten heranziehen, um mit deren Hilfe seine Aufgaben bewältigen zu können, und gleichzeitig gedankliche Anstrengungen weitestgehend zu minimieren (Bär, 2022).

Dies bedeutet, dass zusätzlich zu den bereits geschilderten Wegen, wie Voreingenommenheiten eines Entwicklers in das Verhalten eines KI-Modells einfließen können, der Entwickler, insbesondere im Zustand der kognitiven Erschöpfung, gefährdet ist, mehr und eventuell sogar schwerwiegendere Vorurteile in das Verhalten des KI-Modells einzubeziehen (Bär, 2022).

### **Übermäßiges Selbstvertrauen**

Sollte ein Entwickler ein zu hohes Maß an *Selbstvertrauen* besitzen, so könnte es sein, dass dieser Signale, die auf eine mögliche Verzerrung des KI-Systems hinweisen könnten, ignorieren würde. Dies würde den Grund haben, dass Entwickler mit solch einem Selbstvertrauen nicht glauben würden, dass sie eventuell einen Fehler gemacht haben oder eine Verzerrung in das KI-Modell einbezogen haben. Dementsprechend reagieren sie mit Ablehnung und Ignoranz gegenüber solchen Hinweisen (Bär, 2022).

Durch dieses Unterkapitel wurde verdeutlicht, dass menschliche Urteile und Entscheidungen sich in KI-Systemen widerspiegeln (Gasser, 2019).

Insbesondere zu treffende Designentscheidungen, bewusste sowie unbewusste Voreingenommenheiten, fehlendes Bewusstsein für Vorurteile und ein nicht vorhandenes diverses Entwicklerteam sind die hier primär thematisierten Ursachen dafür, wie Datenwissenschaftler algorithmische Voreingenommenheit in das KI-System bzw. AES einführen.

## **9.5 Komplexität und Transparenz**

Im folgenden Kapitel wird die erhöhte *Komplexität* und die damit zusammenhängende fehlende *Transparenz* von KI-Modellen hervorgehoben. Diese beiden Eigenschaften hängen voneinander ab und liegen aufgrund der Art und Weise, wie KI-Modelle funktionieren und konstruiert sind, vor. Sie wären zwar keine direkte Ursache für algorithmische Voreingenommenheit, jedoch eine dafür, dass algorithmische Voreingenommenheit schwieriger in KI-Systemen zu entdecken sowie zu beheben wäre (Fiegele, 2021; Machill, 2020).

Leistungsstarke und aussagekräftige KI-Modelle basieren oftmals auf neuronalen Netzen. Hier könnte der Entwickler zwar beispielsweise den Rahmen vorgeben, also aus wie vielen Ebenen und Schichten dieses Netzwerk bestehen soll. Nichtsdestotrotz wäre das Netzwerk für die freie Entfaltung des KI-Modells auf dieses entworfen und das KI-Modell würde letzten Endes sein Verhalten anhand des Netzwerkes konstruieren. Also basiert die Entscheidungsfindung des Modells auf diesem neuronalen Netzwerk, sodass man anhand dessen ein erzeugtes Ergebnis

des Modells nachvollziehen könnte. Diese neuronalen Netze und somit die zugrundeliegenden KI-Modelle sind jedoch nicht trivial zu überblicken und zu verstehen. Hierauf wird im Folgenden näher eingegangen (Fiegele, 2021; Machill, 2020).

### **9.5.1 KI-Modelle als Black-Box**

Generell gilt, desto komplexer ein KI-Modell, desto präziser seine Ergebnisse. Man könnte die Komplexität eines KI-Systems unter anderem darauf zurückführen, wie viele Merkmale ein System zur Entscheidungsfindung einbezieht. Denn desto mehr Variablen das System berücksichtigt, desto komplexer wird es (Fiegele, 2021; Machill, 2020).

Die Komplexität kann Maße annehmen, bei welchen das KI-Modell von den Endnutzern oder sogar von den Entwicklern selbst als *Black-Box* angesehen wird. Insbesondere für die Endnutzer könnte es so wirken, als ob sie dem Modell Eingabedaten übergeben und dieses dann eine Vorhersage bzw. Entscheidung zurückliefert, ohne dass diese Nutzer verstehen, wie das System zu genau diesem Resultat gekommen ist (Fiegele, 2021; Machill, 2020).

Die Komplexität des Systems wird zusätzlich durch seinen selbstlernenden Charakter, also durch die bereits vorgestellten „Feedback-Schleifen“ erhöht und würde die Unergründlichkeit der Ergebnisse noch weiter verstärken (Fiegele, 2021; Gasser, 2019; Machill, 2020).

Bezieht man dieses Problem auf AES, die über Menschen urteilen und Menschleben potenziell stark beeinflussen können, so liegt es nahe, dass die Endnutzer bzw. diejenigen, die die Entscheidungen des AES autonom durchführen lassen oder die Entscheidungen als Handlungsempfehlung heranziehen, ein tiefes und ausführliches Wissen über die Verhaltensstruktur des AES besitzen sollten, damit diese bewerten können, ob das KI-System in seiner Entscheidungsfindung korrekt vorgeht oder ob eventuell Voreingenommenheiten einbezogen werden. Aufgrund der erhöhten Komplexität eines KI-Modells bzw. AES und der damit verbundenen Undurchschaubarkeit des Verhaltens und somit der Entscheidungsfindung, wird die Entdeckung von Fehlern und Vorurteilen innerhalb des Systems erschwert. Auf diese *Intransparenz* wird im nächsten Schritt näher eingegangen (Fiegele, 2021; Machill, 2020).

### **9.5.2 Intransparenz von KI-Modellen**

Wie mit dem vorherigen Abschnitt verdeutlicht, können KI-Modelle sehr undurchschaubar werden, sodass das erzeugte Ergebnis des Modells für die Endnutzer oder sogar die Entwickler des Systems nicht nachvollziehbar wäre (Fiegele, 2021; Machill, 2020).

Menschen, die von einem AES eine Absage beispielsweise zu einer Jobbewerbung erhalten, würden diese Entscheidung nicht so stark tolerieren wie eine Absage, die sie von einem anderen Menschen, beispielsweise einem Recruiter, erhalten. Dies hat den Grund, dass diese Bewerber sich vom AES ungerecht behandelt fühlen und die Annahme treffen könnten, dass das über sie urteilende System ihre Stärken und besten Eigenschaften übersehen hat bzw. nicht in die Entscheidungsfindung miteinbezogen hat (Fiegele, 2021; Machill, 2020).

Wenn das AES eine Erläuterung der Entscheidungsfindung zusätzlich zur Entscheidung mitliefern würde, so könnten Menschen, die diesen KI-Modellen eher weniger vertrauen, anhand dieser Erklärung eventuell eine bessere Vorstellung davon bekommen, wie das System konkret zu seiner Entscheidung gelangt ist und ob sie wirklich ungerecht behandelt wurden. Dies könnte das Vertrauen der Menschen gegenüber autonom entscheidende AES verstärken (Fiegele, 2021; Machill, 2020).

Aufgrund der Tatsache, dass diese AES so komplex und sogar für den Entwickler intransparent sowie nahezu unergründlich sein können, kommt es dazu, dass solch eine Erläuterung mit einem erhöhten Aufwand verbunden ist, da die Entwickler hierfür selbst eventuell erst einmal das Verhalten des KI-Systems verstehen müssten und dementsprechend ist dies noch kein Unterfangen, welches standardmäßig in AES wiederzufinden wäre (Fiegele, 2021; Machill, 2020).

Sollte der vom System Betroffene darauf beharren, eine Begründung für die vom System getroffene Entscheidung zu erhalten, so wäre dies wahrscheinlich mit unvermeidbaren finanziellen Kosten verbunden, da es an dieser Stelle nicht ungewöhnlich ist, dass ein Betroffener einen Anwalt hinzuzieht. Selbst wenn der Betroffene letzten Endes Einsicht in das Verhalten des KI-Modells gewährt bekommt, müsste er dieses auch erst einmal verstehen (Fiegele, 2021; Machill, 2020).

Abgesehen von all dem sind sich Betroffene oftmals gar nicht bewusst darüber, dass ein KI-Modell über sie geurteilt hat und sie eventuell sogar von diesem diskriminiert wurden (Fiegele, 2021; Machill, 2020).

Hiermit wird verdeutlicht, welche Ohnmacht und Unwissenheit die Betroffenen aufgrund der Intransparenz dieser KI-Systeme erfahren. Also kann ein AES eventuell diskriminierend über sie urteilen und entweder bemerken die Betroffenen nicht, dass sie algorithmisch diskriminiert worden sind, oder sie bemerken es, jedoch wäre die Zugänglichkeit zu dem zugrundeliegenden Verhalten des KI-Systems mit hohem Aufwand sowie hohen Kosten verbunden und oftmals ist rechtliche Unterstützung von Nöten (Fiegele, 2021; Machill, 2020).

Deutlich wird, dass die erhöhte Komplexität von AES sowohl Vor- als auch Nachteile mit sich bringt. Die erhöhte Komplexität würde zu genaueren Ergebnissen des AES führen. Nichtsdestotrotz würde die mit der erhöhten Komplexität einhergehende zunehmende Intransparenz des Systems letzten Endes dafür sorgen, dass algorithmische Voreingenommenheiten und daraus resultierende diskriminierende Entscheidungen für die Entwickler, Endnutzer und die Betroffenen schwieriger zu entdecken sowie für die Entwickler schwerer zu beheben wären. Folglich bleiben algorithmische Voreingenommenheiten im System bestehen und würden aufgrund von „Feedback-Schleifen“ und dem damit verbundenen eigenständigen Lernen des Systems, sich weiter in das Verhalten des Modells verfestigen (Fiegele, 2021; Gasser, 2019; Machill, 2020).

## **9.6 Benutzerfehler**

Im KI-Algorithmus aufkommende Fehler oder Voreingenommenheiten sind nicht zwangsläufig auf die Daten oder Entwickler des Systems zurückzuführen. Wie der Endnutzer das System verwendet oder wie dieser die Ergebnisse des Systems interpretiert, kann ebenfalls eine mögliche Ursache von Fehlverhalten oder algorithmischer Verzerrung sein.

Auf diese Ursachen wird im Folgenden näher eingegangen.

### 9.6.1 **Übertragung von Kontextverzerrungen**

Angefangen mit der *Übertragung von Kontextverzerrungen*. Angenommen, man hätte ein KI-System, welches für den initialen Kontext, für welchen es entwickelt worden ist, optimal funktioniert und weder Voreingenommenheit noch Inkorrektheit aufweist (Danks & London, 2017; Mohammad, 2021).

Würde der Nutzer sich nun dazu entscheiden, dieses System in einem anderen Kontext zu verwenden, so könnten dort Fehler oder Vorurteile entstehen. Diese wären dann eben kein Resultat des Systems bzw. Seiner Entwicklung, sondern diese kommen durch die falsche und aus dem Kontext gerissene Verwendung des Systems zum Vorschein (Danks & London, 2017; Mohammad, 2021).

Entwickler gestalten das System und wählen Trainings- sowie Testdaten dem Kontext der Verwendung des KI-Modells entsprechend. Die Erzeuger dieser KI-Systeme bzw. AES können die problemfreie Nutzung des Systems dementsprechend maximal für den jeweiligen initial geschilderten Kontext versprechen, da die Entwickler das Modell auf Grundlage dieses Kontextes erstellen (Danks & London, 2017; Mohammad, 2021).

Also wären die hier aufkommenden Verzerrungen und Fehler weder Schuld der Entwickler noch der übergebenen Trainingsdaten, sondern die der Endnutzer (Danks & London, 2017; Mohammad, 2021).

Dies lässt sich anhand eines Beispiels verdeutlichen.

Man gehe davon aus, dass Entwickler ein KI-System entworfen haben, welches dazu verwendet wird, autonomes Fahren von PKWs in den USA zu ermöglichen. Weiterführend gehe man davon aus, dass es fehlerfrei und korrekt arbeitet (Danks & London, 2017; Mohammad, 2021).

Sollte der Nutzer des Systems sich nun dazu entscheiden, dieses in Großbritannien einzusetzen, so würde es nichtmehr so optimal wie in den USA funktionieren, da in Großbritannien teilweise andere Verkehrsregeln gelten als in den USA. Die größte Änderung wäre, dass man links fährt anstatt wie in den USA oder auch Deutschland rechts zu fahren (Danks & London, 2017; Mohammad, 2021).



Sollte das KI-System nicht das alte Verhalten verlernt und das neue, auf Großbritannien fokussierte, Verhalten erlernt haben, so würde es in seiner Verwendung in Großbritannien Fehlverhalten aufweisen. Dies hätte potenziell sehr schwerwiegende Folgen, wie beispielsweise mögliche resultierende Autounfälle (Danks & London, 2017; Mohammad, 2021).

Dieses Beispiel illustriert mögliche negative Auswirkungen, wenn ein Nutzer den KI-Algorithmus außerhalb seines Anwendungskontextes, für welchen er entwickelt und trainiert worden ist, verwendet. Hiermit wird zwar keine Person diskriminiert, jedoch hätte der veränderte Anwendungskontext des Systems fatale Folgen auf dessen tatsächliche Nutzung (Danks & London, 2017; Mohammad, 2021).

### 9.6.2 Verzerrung durch Missinterpretation

Voreingenommenheiten können auch entstehen, wenn die Ergebnisse des AES vom Nutzer missinterpretiert werden (Swaminathan, 2022).

Diese mögliche Ursache von Verzerrung entsteht also aus der Fehlinterpretation der hervorgebrachten Resultate eines KI-Systems. Dies geschieht entweder wegen der Tatsache, dass dem Nutzer das Verständnis für die Funktionsweise dieses Systems fehlt und er somit nicht einschätzen kann, was mit dem jeweiligen Ergebnis gelöst oder beantwortet werden soll oder, angenommen die Ergebnisse eines KI-Modells werden als Input für einen anderen Algorithmus genutzt, der Nutzer hat eine falsche Vorstellung davon, welchen Output er benötigt, um den Anforderungen des Inputs des anderen Algorithmus gerecht zu werden und interpretiert die Ergebnisse des Output-liefernden Modells somit falsch (Mohammad, 2021).

Eine weitere Ursache für die Fehlinterpretation des Nutzers könnte sein *mangelndes Hintergrundwissen* bezüglich des Anwendungsfeldes des Systems sein (Danks & London, 2017; Machill, 2020).

Diese mögliche Ursache für algorithmische Voreingenommenheit lässt sich anhand eines Beispiels näherbringen.

Angenommen, man hätte ein KI-System, welches vorhersagen soll, wie viele der Männer und Frauen, die sich für einen Studiengang an einer Universität beworben haben, von der Uni dann auch tatsächlich akzeptiert werden (Machill, 2020).

Beispielhaft könnte das KI-System die Vorhersage treffen, dass von der Universität rund 45% der Männer und rund 25% der Frauen akzeptiert werden. Der Nutzer dieses KI-Systems könnte dieses Ergebnis nun als Indikator dafür identifizieren, dass dieses KI-Modell geschlechtvoreingenommen und diskriminierend gegenüber Frauen ist (Machill, 2020).

Nun ist es jedoch so, dass diese Annahme des Nutzers eine Fehlinterpretation des Ergebnisses illustriert. Diese ist möglicherweise durch fehlendes Hintergrundwissen verursacht worden. Denn berücksichtigt man die Hintergrundinformationen, so wird deutlich, wieso Frauen eine niedrigere Akzeptanzquote als Männer aufweisen (Machill, 2020).

Angenommen, diese Universität weist sechs Departements auf und das KI-System liefert entweder nur die zusammengerechneten Akzeptanzquoten der Geschlechter oder die Quoten der Geschlechter je Departement zurück. Betrachtet man die Akzeptanzquote von den einzelnen Departements, so weisen Frauen in vier Departements eine höhere Akzeptanzrate als Männer auf. Weiterführend würden sie bei den anderen beiden Departments mit den Quoten der Männer gleichauf sein (Machill, 2020).

Nun stellt sich jedoch die Frage, wieso die gesamte Akzeptanzquote der Frauen so niedrig liegt. Dies hat den Grund, dass Männer sich vielmehr bei Fakultäten beworben haben, welche eher eine größere Menge an Bewerbern akzeptieren, während Frauen sich vielmehr bei Fakultäten beworben haben, welche eher weniger Bewerber akzeptieren (Machill, 2020).

Auf Grundlage dieses Hintergrundwissens wäre der vom Nutzer getätigte Vorwurf, dass dieses KI-System voreingenommen gegenüber Frauen ist, widerlegt (Machill, 2020).

Mit diesem Beispiel wurde illustriert, wie Nutzer ein Ergebnis insofern missinterpretieren können, dass somit einem KI-Modell fehlerhaftes- oder voreingenommenes Verhalten zugeschrieben wird, obwohl dieses optimal und neutral funktioniert (Danks & London, 2017).

In diesem Kapitel wurde beispielhaft illustriert, wie bei einem eigentlich unvoreingenommenen und korrekt arbeitenden KI-System aufgrund der vorgestellten Benutzerfehler trotzdem

Fehlverhalten und algorithmische Voreingenommenheit hervorkommt oder irrtümlich erkannt wird.

## **9.7 Rechtliche Regelungen**

In diesem Kapitel wird es einen Überblick über die deutschen- und internationalen Menschen- sowie Grundrechte geben. Weiterführend werden auf die rechtlichen Regulierungen von künstlicher Intelligenz und insbesondere AES eingegangen. Abschließend wird verdeutlicht, wie trotz dieser Regelungen, algorithmische Voreingenommenheit noch so präsent sein kann.

Die bestehenden Diskriminierungsrisiken, die mit dem Einsatz von KI-Systemen verbunden sind, wurden in dieser Arbeit ausgiebig erläutert. Demensprechend liegen einem die möglicherweise schwerwiegenden Folgen der Nutzung und Integrierung dieser Systeme in ein Unternehmen oder eine Staatseinheit vor Augen.

Es stellt sich heraus, dass regulierend einzugreifen zunehmend relevanter und dringlicher wird. Somit wären kürzliche Ereignisse wie ein im Jahr 2021 von der EU-Kommission veröffentlichter „[...] Verordnungsentwurf für die Regulierung von KI [...]“ (Öktem, 2022, S. 1), ein im Oktober 2020 in Deutschland verabschiedeter Abschlussbericht zu den Chancen sowie Hürden von KI oder eine im Dezember 2020 in Deutschland weitergeschriebene KI-Strategie, welche die Nutzung von KI im Land fördern soll, wichtige Schritte in eine regulierte und somit weniger algorithmisch voreingenommene KI-Zukunft (Öktem, 2022).

Durch den Einsatz von KI-Systemen bzw. AES steht die Verletzung des Rechts auf Privatsphäre und des Diskriminierungsverbotes in großer Gefahr. Deren Verletzung wurde mit den vielen Fallbeispielen illustriert und zieht oftmals weitere Menschenrechtsverletzungen nach sich (Öktem, 2022).

Unter anderem ist „[...] das Recht auf Zugang zu Abhilfe von besonderer Bedeutung [...]“ (Öktem, 2022, S. 2). Denn wenn jemand von den Entscheidungen eines AES betroffen ist oder ihm bewusst wird, dass er möglicherweise diskriminiert wurde, so wird es wie im Unterkapitel „Intransparenz von KI-Modellen“ bereits vorgestellt, nicht trivial sein, Einsicht in die zugrundeliegende Verhaltensweise des AES zu erlangen (Öktem, 2022).

Allerdings sind Staaten aufgrund der Achtungspflicht dazu verpflichtet, die Menschenrechte eines Bürgers „[...] zu achten, zu schützen und zu gewährleisten“ (Öktem, 2022, S. 4). Dies bedeutet, wenn staatliche Einheiten AES verwenden, die Menschen in ihren Menschenrechten nicht verletzt werden dürfen. Ein Beispiel für das Nichteinhalten dieser Verpflichtung konnte mit der vorgestellten staatlichen Verwendung des AMAS-KI-Systems näher erläutert werden (Öktem, 2022).

Staaten sind zusätzlich dazu verpflichtet, den Menschen die Ausübung ihrer Menschenrechte zu ermöglichen. Demensprechend hat der Staat ebenfalls die Verpflichtung, seine Bewohner vor Menschenrechtsverstößen durch außerstaatliche Einheiten wie Unternehmen zu schützen. Dementsprechend haben Staaten die Pflicht regulierend zu intervenieren, wenn Unternehmen beispielsweise nach Geschlecht diskriminieren. Hierbei ist egal, ob diese Diskriminierung von Menschen oder Algorithmen oder beiden durchgeführt wird. Solche vom Unternehmen ausgehende Diskriminierung wurde bereits mit vielen Beispielen, unter anderem mit dem Amazon-Fallbeispiel, illustriert (Öktem, 2022).

Vor diesem Hintergrund wird im nächsten Schritt die deutsche rechtliche Situation bezüglich algorithmischer Voreingenommenheit betrachtet.

Hier wäre das *Grundgesetz (GG)*, das *Allgemeine Gleichbehandlungsgesetz (AGG)* sowie die *Datenschutz-Grundverordnung (DSGVO)* von besonderer Bedeutung (Kolleck & Orwat, 2020).

Welche geschützten Merkmale diese Gesetze sichern, wird mit folgender Abbildung erkenntlich (Kolleck & Orwat, 2020).

Geschütztes Merkmal	GG	AGG	DSGVO
Ethnie	✓	✓	✓
Herkunft	✓		
Geschlecht	✓	✓	
Sprache	✓		
Politische Meinung	✓		✓
Weltanschauung	✓	✓	✓
Behinderung	✓	✓	
Alter		✓	
Gesundheitszustand	✓		✓
Sexuelle Orientierung		✓	✓

Abbildung 13: Geschützte Merkmale laut Deutschem Gesetz (eigene Darstellung) (Kolleck & Orwat, 2020)

In Artikel 3 des GG wird hervorgehoben, dass alle Menschen gleichbehandelt werden sollen und insbesondere nicht aufgrund ihrer geschützten Merkmale benachteiligt werden dürfen. Weiterführend wäre die Wahrung der Menschenrechte in diesem Gesetzesbuch niedergelegt (*Grundgesetz für die Bundesrepublik Deutschland Art 3, o. D.; Kolleck & Orwat, 2020; Schutz von Menschenrechten, o. D.*).

Einige der im GG angesprochenen geschützten Merkmale finden sich auch in den Zielen des AGG unter §1 wieder (*Allgemeines Gleichbehandlungsgesetz (AGG) § 1 Ziel des Gesetzes, o. D.; Kolleck & Orwat, 2020*).

Unter Berücksichtigung der Gesetze des AGG wäre die Verwendung von geschützten Merkmalen „[...] bei Differenzierungsentscheidungen mithilfe von Algorithmen bzw. Computersystemen unzulässig“ (Kolleck & Orwat, 2020, S. 40). Weiterführend wäre, wie in §3 AGG beschrieben, die Verwendung der bereits vorgestellten Proxy-Variablen ebenfalls unerlaubt,

außer die Nutzung wäre rechtmäßig gerechtfertigt (*Allgemeines Gleichbehandlungsgesetz (AGG) § 1 Ziel des Gesetzes*, o. D.; Kolleck & Orwat, 2020).

Weiterführend würde der Erwägungsgrund 71 der DSGVO es staatlichen Einheiten sowie Unternehmen untersagen, das diese AES autonom und ohne menschliche Entscheidungsentität Urteile über Personen treffen, die rechtliche Auswirkungen oder andere ähnlich beeinträchtigende Auswirkungen auf diese betroffenen Personen haben. Hierzu würden beispielsweise automatische algorithmische Ablehnungen von Bewerbungsverfahren gelten (*Erwägungsgrund 71 - Profiling*, o. D.; Kolleck & Orwat, 2020; *Verarbeitung besonderer Kategorien personenbezogener Daten*, o. D.).

Mit diesen Regelungen und deren Erläuterung wurde illustriert, wie Deutschland die geschützten Merkmale eines Menschen absichern möchte und weiterführend die Menschenrechte versucht zu wahren.

Nichtsdestotrotz sind die Menschenrechte und folglich die geschützten Merkmale nicht nur in Gesetzen des eigenen Landes, sondern ebenfalls in internationalen Regelungen fest verankert (Öktem, 2022).

Die im Jahr 1948 von den Vereinten Nationen (UN) veröffentlichte *Allgemeine Erklärung der Menschenrechte* war die Grundlage für ein umfangreiches Vertragskonstrukt der UN, welches die Rechte eines Menschen in etlichen Verträgen schriftlich niedergelegt hat, sodass diese auf internationaler Ebene ebenfalls ganzheitlich abgedeckt und ein fester Bestandteil sind (*Allgemeine Erklärung der Menschenrechte*, 2023; Öktem, 2022).

Zu diesen Verträgen gehören insbesondere „[...] der Zivilpakt, der Sozialpakt, die Anti-Rassismus-Konvention, die Frauenrechtskonvention, die Anti-Folter-Konvention, die Kinderrechtskonvention, die Behindertenrechtskonvention [...]“ (*Schutz von Menschenrechten*, o. D.). Diese sind speziell mit Blick auf „[...] Diskriminierungsverbote und Teilhaberechte [...]“ (Öktem, 2022, S. 4) von Relevanz (Öktem, 2022; *Schutz von Menschenrechten*, o. D.).

Zu denjenigen Ländern, welche vielen oder sogar allen UN-Menschenrechtsabkommen zugestimmt und sich diesen verpflichtet haben, gehören fast alle Mitgliedstaaten der UN. Hierunter auch Deutschland, Österreich oder die USA. Also all diejenigen Länder, welche in den

vorgestellten Fallbeispielen Verwendung finden, sind die Verpflichtung eingegangen, diese Menschenrechte zu wahren (Höger, 2023; *Human Rights*, 2023).

Wichtig an dieser Stelle hervorzuheben wäre, dass definitiv in den etlichen menschenrechtsbewahrenden Verträgen der UN und eventuell ebenfalls in gewissen deutschen Gesetzesregelungen nicht speziell auf Rechte im digitalen Raum eingegangen wird. Nichtsdestotrotz liegt den Menschenrechten genauso wie den Grundrechten „[...] ein dynamisches Verständnis zugrunde [...]“ (Öktem, 2022, S. 4). Dies würde bedeuten, dass die Auslegung der Menschen- und Grundrechte einerseits an jeglichen gesellschaftlichen und technologischen Weiterentwicklungen ausgerichtet ist und sich andererseits an der jeweiligen Lebenssituation orientiert. Dementsprechend entfalten alle im internationalen Raum gesicherten Menschen- sowie Grundrechte „[...] ihre Wirkung deshalb offline wie online [...]“ (Öktem, 2022, S. 4) (Öktem, 2022).

Zusätzlich zur dynamischen Auslegung der Menschenrechte deutet der UN-Menschenrechtsrat im gleichen Sinne darauf hin, die Menschenrechte ebenso auf digitaler Ebene zu bewahren (Öktem, 2022).

Dies würde bedeuten, dass auf internationaler Ebene all diejenigen Länder und zugrundeliegenden staatlichen Einheiten und Unternehmen, welche AES einsetzen, ebenfalls auf die menschenrechtsbewahrende Nutzung dieser zu achten haben, falls sie sich dazu verpflichtet haben, einige oder alle UN-Menschenrechtsverträge anzuerkennen und einzuhalten.

Nichtsdestotrotz würden viele solcher Länder, wie mit den Fallbeispielen vorgestellt, trotzdem ihre Bürger mittels staatlicher Einheiten algorithmisch diskriminieren oder von Unternehmen algorithmisch diskriminieren lassen.

Nun stellt sich jedoch die Frage, wieso trotz dieser umfassenden rechtlichen Grundlagen trotzdem noch viele AES existieren, die die Menschenrechte und insbesondere die geschützten Merkmale eines Menschen verletzen und somit bestimmte Personen, insbesondere Minderheiten, diskriminieren.

Zur Beantwortung dieser Frage kann man Bezug zu den bereits erläuterten Ursachen für das Entstehen oder Bestehenbleiben von algorithmischer Voreingenommenheit nehmen.

Angefangen mit unzureichendem Überwachen sowie Bewusstsein für algorithmische Voreingenommenheit innerhalb des Entwicklungsprozesses. Die etlichen Ursachen illustrieren, wie unterschiedliche Voreingenommenheiten sich zu verschiedenen Zeitpunkten in das KI-System bzw. AES einschleusen können. Dementsprechend muss das Entwicklerteam im Grunde jederzeit damit rechnen, dass mögliche Verzerrungen sich in Form von algorithmischer Voreingenommenheit in das System begeben und das Verhalten des Systems beeinflussen können (De Montis, 2019).

Also sollte das Entwicklerteam kein Bewusstsein für algorithmische Verzerrung haben und weiterführend keine gründliche sowie regelmäßige Überwachung ihres KI-Systems bzw. AES vornehmen, so ist die Wahrscheinlichkeit hoch, dass sich Vorurteile unbemerkt ins System einschleusen und eventuell sogar während der Nutzung des Systems unbemerkt bleiben (De Montis, 2019).

Die Annahme, dass diese Voreingenommenheiten eventuell sogar während der Verwendung des Systems unbemerkt bleiben, hat den Grund, dass diejenigen Systeme, welche aussagekräftige Ergebnisse liefern, oftmals auf einem neuronalen Netzwerk beruhen, welches diese Systeme sehr komplex und intransparent machen kann. Weiterführend wäre bei Verdacht einer Voreingenommenheit somit noch die Hürde übrig, die Ursache im Netzwerk zu lokalisieren und zu beheben, sodass im schlimmsten Fall der Entwickler eine tatsächlich bestehende Voreingenommenheit entweder nicht entdeckt und folglich davon ausgeht, dass keine existiert und das System korrekt sowie unvoreingenommen arbeitet oder dass die Ursache zwar entdeckt wird, das Netzwerk jedoch so komplex und intransparent ist, dass eine Behebung kaum möglich ist.

Diese Komplexität kommt ebenfalls durch die Fähigkeit des AES hervor, belanglose Zusammenhänge in Daten zu erkennen. Dies lässt sich beispielhaft näherbringen.

Angenommen, ein auf neuronalem Netzwerk basierendes AES, welches in den USA Anwendung finden soll, würde unter anderem die Merkmale „Hautfarbe“, „Postleitzahl“ sowie „X“ berücksichtigen. Wie in einem früheren Kapitel erörtert, kann mithilfe der „Postleitzahl“ auf die „Hautfarbe“ einer Person geschlossen werden. Das Merkmal „X“ stellt in diesem Beispiel eine mit der Hautfarbe bedeutungslos zusammenhängende Variable dar.



Sollten die Entwickler dieses Systems nun zur Vermeidung von algorithmischer Voreingenommenheit das geschützte Merkmal „Hautfarbe“ sowie die damit zusammenhängende Variable „Postleitzahl“ außen vor lassen, so könnte das AES aufgrund seines zugrundeliegenden Lernverhaltens und seines neuronalen Netzwerkes das bedeutungslos zusammenhängende Merkmal „X“ heranziehen, um daraus auf das geschützte Attribut der „Hautfarbe“ zu schließen. Hierbei wichtig zu erwähnen wäre, dass die Entwickler das Merkmal „X“ nicht ebenfalls aus dem AES entfernt haben, da sie nicht damit gerechnet hatten, dass sich das System mit diesem willkürlichen Attribut einen Zusammenhang zur „Hautfarbe“ erschließt.

Hieraus resultierende algorithmische Voreingenommenheit bzw. diskriminierende Entscheidungen gegenüber dunkelhäutigen Personen wären in solch einem System schwieriger zu ermitteln als in einem System, welches geschützte Merkmale oder zumindest mit diesen Merkmalen tatsächlich zusammenhängende Attribute direkt berücksichtigt.

Dies illustriert, dass das in den deutschen Gesetzen explizit geforderte Außenvorlassen von geschützten sowie damit korrelierenden Merkmalen aufgrund des komplexen Lernverhaltens solcher AES darin münden kann, dass die entfernten geschützten Merkmale trotzdem über Platzhalter von Platzhaltern zur Entscheidungsfindung des AES herangezogen werden können. Dementsprechend kann einerseits weiterhin algorithmische Voreingenommenheit aufkommen und andererseits wäre diese Voreingenommenheit schwieriger im System zu ermitteln sowie zu entfernen, da nicht logisch darauf zurückgeführt werden kann, wie das System sich die geschützten Merkmale abgeleitet hat.

Zusätzlich zu gesetzlichen Regelungen, die eventuell kontraproduktiv sind, finden die internationalen sowie deutschen Regelungen häufig nur aufgrund ihrer dynamischen Interpretierbarkeit auf algorithmische Voreingenommenheit Anwendung. Der fehlende spezielle Bezug der gesetzlichen Regelungen auf algorithmische Voreingenommenheit von AES könnte darin münden, dass Voreingenommenheiten eventuell nicht erfasst werden, da diese Regelungen dafür möglicherweise nicht ausführlich und detailliert genug formuliert sind oder dass die Durchsetzung von Sanktionen gegenüber denjenigen AES, welche nachgewiesen diskriminierend sind, mit der aktuellen Rechtsgrundlage ebenfalls eine Schwierigkeit darstellt, da unter anderem die Schwere der Auswirkungen von algorithmischer Voreingenommenheit stark variiert (Kolleck & Orwat, 2020).

Aus diesem Grund hat die EU-Kommission im Jahr 2021 einen „[...]“ Vorschlag für ein Gesetz über Künstliche Intelligenz [...]“ (*Europäisches Gesetz über Künstliche Intelligenz*, o. D.) veröffentlicht, auf welchen in einem späteren Kapitel näher eingegangen wird (*Europäisches Gesetz über Künstliche Intelligenz*, o. D.).

In diesem Unterkapitel konnte im Kern die aktuelle deutsche und internationale rechtliche Situation bezüglich algorithmischer Voreingenommenheit von AES nähergebracht werden. Weiterführend wurde erläutert, wieso trotz dieser Regelungen heutzutage noch immer algorithmisch voreingenommene Systeme existieren.

Mit diesem gesamten Kapitel wurden, wie angekündigt, die relevanten Ursachen für das Entstehen sowie das Bestehenbleiben von algorithmischer Voreingenommenheit detailliert erläutert. Darüber hinaus wurden die rechtlichen bzw. regulatorischen Mängel bezüglich dieses Themas veranschaulicht und damit verdeutlicht, wie aufgrund dieser noch immer algorithmisch voreingenommene AES eingesetzt werden. Im Gesamten wurde ein allumfassendes Bild bezüglich der Thematik, wie algorithmische Voreingenommenheit aufkommen und sich verfestigen kann vorgestellt.

Letztendlich hebt dieses Kapitel, zusammen mit den Fallbeispielen, die Notwendigkeit der Behebung bzw. der Minderung dieser Ursachen bzw. der algorithmischen Voreingenommenheiten hervor.

## 10 Identifizierung & Minderung von algorithmischer Voreingenommenheit

Im Folgenden wird erläutert, wie die Ursachen für algorithmische Voreingenommenheit identifiziert und gemindert werden können, um ein weniger verzerrtes KI-System vorliegen zu haben.

Hierbei werden insbesondere die menschlichen, datentechnischen sowie rechtlichen Aspekte thematisiert. Weiterführend werden unterstützende Tools vorgestellt und es wird die Relevanz von Überwachung sowie Überprüfung dieser Systeme hervorgehoben.

Bevor das erste Unterkapitel vorgestellt wird, ist es relevant zu wissen, dass bislang und im weiteren Verlauf dieser Arbeit von *algorithmische Voreingenommenheit mindernden Maßnahmen* die Rede ist und war. Dies hat den Grund, dass - dem Autor dieser Arbeit nach - durchführbare Maßnahmen, um algorithmische Voreingenommenheit zu bekämpfen, keine Garantie dafür sind, dass KI-Systeme nach Durchführung dieser Maßnahmen wirklich komplett unvoreingenommen arbeiten.

Diese Aussage begründet der Autor einerseits mit den vielen verschiedenen Möglichkeiten, wie Vorurteile sich in das KI-System einschleusen können, und andererseits damit, dass diese Systeme so komplex sind, dass wahrscheinlich nicht alle Voreingenommenheiten aufgedeckt werden können und dementsprechend im KI-System bzw. AES fest verankert bleiben.

## **10.1 Menschliche Maßnahmen**

Zuerst werden Maßnahmen vorgestellt, mit welchen menschliche Verzerrungen identifiziert und mitigiert werden können.

Dieses Kapitel beginnt mit den menschlichen Voreingenommenheiten, da diese, wie in Abbildung 11 veranschaulicht, der Ursprung aller algorithmischen Voreingenommenheiten sind.

Somit sieht man die Dringlichkeit, als Erstes die menschlichen Ursachen für algorithmische Verzerrung zu mindern, damit die Grundlage geschaffen wird, weitere Mitigierungsmaßnahmen wirkungsvoll durchführen zu können.

### **10.1.1 Diversität**

Bereits bestehende Voreingenommenheiten eines Menschen unterliegen seinen kognitiven Verzerrungen und somit seinen individuellen Eigenschaften, seinem Lebensverlauf sowie den dort aufgetretenen Erlebnissen. Aufgrund der Tatsache, dass Unterschiede bei geschützten Merkmalen wie z. B. Geschlecht oder Ethnie, Auswirkungen darauf haben können, welche Erlebnisse und welchen Lebensverlauf man als Mensch hinterlegt, liegt es nahe, dass sich somit ebenfalls unterschiedliche Voreingenommenheiten bei den verschiedenen Menschen bilden und verfestigen (Gasser, 2019; Machill, 2020; Marques, 2021).

Berücksichtigt man nun die in der KI-Branche vorherrschende homogene Gruppe an hellhäutigen männlichen Entwicklern, so liegt einem KI-entwickelndem Unternehmen wahrscheinlich kein vielfältiges Entwicklerteam zugrunde, wodurch eventuell ebenfalls wenige gravierende Differenzen zwischen den individuellen Lebenshintergründen vorliegen. Dementsprechend könnte solch ein Entwicklerteam ähnliche Weltauffassungen und Voreingenommenheiten teilen, wodurch weniger Bewusstsein für diese Voreingenommenheiten existiert. Folglich werden andere Personengruppen, z. B. Frauen oder dunkelhäutige Individuen letzten Endes von den entworfenen KI-Systemen diskriminiert, da die Voreingenommenheiten gegenüber diesen Personengruppen unbehandelt blieben und sich somit in das Verhalten des Systems einschleusen konnten (Gasser, 2019; Machill, 2020; Marques, 2021).

Dementsprechend kommt die Notwendigkeit von *Diversität* innerhalb der KI-Branche, genauer, die Dringlichkeit von vielfältigeren KI-Entwicklerteams, hervor. Es wird empfohlen, zunehmend Minderheiten wie beispielsweise dunkelhäutige Personen in der KI-Branche, insbesondere für Führungsrollen, einzustellen (Machill, 2020).

Solch ein diverseres KI-Entwicklerteam würde also Personen beinhalten, welche aktuell am ehesten von KI-Systemen diskriminiert werden. Die Hinzunahme solcher Entwickler würde die Fähigkeiten und Ansichten des Teams komplementieren und dafür sorgen, dass das Entwicklerteam unvoreingenommener arbeitet. Unter anderem aus dem Grund, dass diejenigen, die Voreingenommenheiten gegenüber einer Personengruppe am frühesten aufdecken, oftmals diejenigen sind, die von diesen Voreingenommenheiten betroffen sind (Marques, 2021).

Sind diese Voreingenommenheiten erstmal aufgedeckt, so können mindernde oder vorbeugende Maßnahmen vorgenommen werden (Marques, 2021).

Dadurch, dass KI-Systeme dazu neigen, Minderheiten zu diskriminieren, wäre die Einstellung von Arbeitskräften, die einer Minderheit zugehörig sind, innerhalb dieser Branche eine Möglichkeit, algorithmische Voreingenommenheit zu mindern (Marques, 2021).

Um als Unternehmen die Vorteile eines diversen KI-Entwicklerteams in Anspruch nehmen zu können, bedarf es einer Unternehmenskultur, in welcher jeder Entwickler seine Ansichten, Gedanken sowie Sorgen teilen kann. Denn nur so würden sich die unterschiedlichen Entwickler komplementieren und nur so kann Aufmerksamkeit sowie Bewusstsein für verzerrte Merkmale, Voreingenommenheiten der Entwickler und Verzerrung der Daten, zum Vorschein gelangen (De Montis, 2019; Gasser, 2019).

Zusätzlich zu einem vielfältigen KI-Entwicklerteam könnte das Unternehmen zukünftige Nutzer sowie Betroffene des KI-Systems bzw. AES ebenfalls in deren Entwicklung einbeziehen, um so zusätzlichen Austausch und damit verbunden, weitere Ansichten sowie Denkanstöße zu erhalten, welche verstärkt Licht auf verschiedenste Voreingenommenheiten werfen würden. Dies könnte darin münden, dass die KI-Entwickler Bewusstsein für diese weiteren Verzerrungen entwickeln und auf dieser Grundlage dann eventuell das System explizit so gestalten, dass diejenigen Voreingenommenheiten, welche ihnen nun bewusst sind, nicht in den Lernprozess des AES einbezogen werden, damit diejenigen Betroffenen, über welche das AES urteilt und

welche eventuell aufgrund dieser Voreingenommenheiten diskriminiert worden wären, im fertigen AES nicht diskriminiert werden, sondern im besten Falle rein objektiv über sie geurteilt wird (Öktem, 2022; Marques, 2021).

Außerdem hätte das Einbeziehen zukünftiger Nutzer in die Entwicklung des KI-Systems zusätzlich den Vorteil, dass diese kontinuierlich im Austausch mit den Entwicklern stehen und somit gemeinsam das Bild des zu entwickelnden KI-Systems konkretisieren können, sodass weniger Missverständnisse vorherrschen und das KI-System genau die Funktionsweise erfüllt, die es erfüllen soll. Weiterführend könnte das Einbeziehen zukünftiger Nutzer in die Entwicklung ebenfalls darin münden, dass diese genau nachvollziehen können, für welchen konkreten Kontext das KI-System entworfen wird. Folglich würden sie eventuell weniger dazu neigen, diesen für einen speziellen Kontext entworfenen Algorithmus in einem anderen Kontext zu verwenden, da ihnen möglicherweise bewusst ist, dass dieser speziell für ihren beschriebenen Kontext gestaltet wurde und somit eventuell nicht gleich präzise Vorhersagen in einem anderen Kontext aufweisen würde. Auf diese Weise würde man also dem „Benutzerfehler“ bzw. der „Übertragung von Kontextverzerrungen“ entgegenwirken (De Montis, 2019; Öktem, 2022; Gasser, 2019; Marques, 2021).

Die Diversifizierung des Entwicklerteams sowie die Einbeziehung von Nutzern und Betroffenen in die Entwicklung des Systems wären mögliche Maßnahmen, um voreingenommene Variablen sowie menschliche- und datentechnische Verzerrungen zu entdecken und folglich diesen vorbeugend entgegenwirken zu können.

### 10.1.2 **Bewusstsein**

An dieser Stelle wichtig zu erwähnen ist, dass ein KI-System, welches keine Voreingenommenheiten enthält, eventuell keine nutzbaren Ergebnisse hervorbringen würde. Algorithmische Verzerrung kann somit zu einer akkurateren Entscheidungsfindung des KI-Systems beitragen (Marques, 2021).

Bislang konnte unter algorithmischer Voreingenommenheit ein KI-System bzw. AES verstanden werden, welches diskriminierendes Verhalten aufweist sowie folglich beeinträchtigende

oder bevorzugende Entscheidungen trifft. Also konnte bisher algorithmische Voreingenommenheit mit algorithmischer Diskriminierung gleichgesetzt werden.

In diesem Unterkapitel werden diese Begriffe jedoch differenziert betrachtet.

Man gehe beispielsweise von einem KI-System aus, welches prognostizieren soll, ob eine Person Brustkrebs hat. Solch ein System könnte dahingehend tendieren, Frauen eher ein positives Ergebnis zuzuordnen als Männern. Dementsprechend wäre das System dahingehend voreingenommen, bei Frauen eher zu einem positiven Ergebnis zu tendieren. Dies hat den Grund, dass Frauen eher Opfer von Brustkrebs werden als Männer. Nun stellt sich die Frage, ob man diese Voreingenommenheit im System behalten oder entfernen sollte. Angenommen, das KI-System würde für Frauen sowie Männer trotzdem eine hohe Genauigkeit der Ergebnisse aufweisen, so würde man sich dazu entscheiden, diese Verzerrung im KI-Modell beizubehalten, da diese sich positiv auf die Genauigkeit des Algorithmus auswirkt und dabei keine diskriminierenden Effekte gegenüber einer Personengruppe aufweist (*Brustkrebs (Mammakarzinom)*, 2022; Gasser, 2019; Marques, 2021).

Solche AES, in welchen die geschützten Merkmale nur zur Genauigkeit beitragen und nicht zur Diskriminierung von Personengruppen herangezogen werden, sind laut dem Autor dieser Arbeit auf Grundlage der bisher vorgestellten Fallbeispiele eher weniger anzutreffen.

Zumindest ist dies bei AES der Fall, welche geschützte Merkmale einbeziehen sowie menschenrechtlich relevante Entscheidungen treffen und somit direkt über Menschen urteilen sowie ihr Leben auf Grundlage dieses Urteils potenziell stark beeinflussen.

Aus diesem Grund wird nach diesem Unterkapitel algorithmische Voreingenommenheit wieder mit algorithmischer Diskriminierung gleichgestellt sein, außer es wird explizit differenziert angegeben.

Aufgrund der Tatsache, dass das Verhalten der KI-Modelle von den ihnen zur Verfügung gestellten Daten abhängig ist und diese Daten von den Entwicklern ausgewählt und zum Lernen aufbereitet werden, entsteht die Notwendigkeit, dass diese Entwickler sich darüber bewusst sind, welche Merkmale voreingenommen sind sowie diskriminierendes oder bevorzugendes Verhalten hervorrufen und somit nicht in das KI-System einbezogen werden sollten und welche Variablen zwar verzerrt sind, aber unter Berücksichtigung des jeweiligen Kontextes des

Systems eher zur Genauigkeit des Modells beitragen und dabei kein diskriminierendes oder bevorzugendes Verhalten aufweisen (Gasser, 2019; Marques, 2021).

Das *Bewusstsein* des Entwicklers einerseits für kognitive Verzerrungen und andererseits für letztlich daraus resultierende voreingenommene Daten, welche je nach Kontext des KI-Systems, diesem zum Erlernen eines akkuraten sowie diskriminierungsfreien Verhalten zugeführt werden können oder nicht, wären die Grundlage dafür, dass Entwickler algorithmische Voreingenommenheiten mindern können, da Entwickler mit solch einem Bewusstsein eher dazu neigen, potenziell diskriminierende Merkmale nicht in das KI-System einzubeziehen als Entwickler ohne diese Wahrnehmung. Somit reicht lediglich das Bewusstsein für kognitive- sowie datentechnische Vorurteile und weiterführend algorithmische Verzerrung aus, um Voreingenommenheiten bzw. Diskriminierungen zu mindern (Gasser, 2019; Machill, 2020; Marques, 2021).

Also bedarf es dieses Bewusstseins bei allen KI-Entwicklern, damit zum einen bestehende Vorurteile der Entwickler sowie Diskriminierungen innerhalb der Daten überhaupt erkannt und folglich mitigiert werden können. Sowie zum anderen ist es nötig, damit weitere algorithmische Diskriminierung mindernde Maßnahmen überhaupt greifen können. Dementsprechend ist ein Bewusstsein für algorithmische Voreingenommenheit notwendig, da man nicht über etwas sprechen kann oder sich über etwas Gedanken machen kann, wovon man keine Wahrnehmung hat (Gasser, 2019; Machill, 2020; Marques, 2021).

An dieser Stelle ist ebenfalls wichtig anzusprechen, dass die Nutzer des KI-Systems ebenfalls solch ein Bewusstsein entwickeln sollten (Machill, 2020; Marques, 2021).

Dies hat einerseits den Grund, dass Nutzer eventuell als finale Entscheidungsentität fungieren und somit ihre eigenen, möglicherweise vorherrschenden Voreingenommenheiten reflektieren sollten, und andererseits, damit diese Nutzer die Handlungsvorschläge der KI-Systeme nicht als neutrale Wahrheit ansehen, sondern deren Ergebnisse kritisch betrachten (De Montis, 2019; Gasser, 2019; Machill, 2020; Marques, 2021).

Betroffene sollten ebenfalls ein Bewusstsein für algorithmische Voreingenommenheit entwickeln, damit sie im Falle einer Diskriminierung überhaupt auf diese aufmerksam werden



können. Im Ergebnis könnten sie somit eventuell sogar rechtlich gegen die Entwickler und/oder die Nutzer des Systems vorgehen (De Montis, 2019; Machill, 2020; Marques, 2021).

Letzten Endes wäre die Entwicklung von Bewusstsein bezüglich der Thematik von algorithmischer Voreingenommenheit ein notwendiger erster Schritt für Entwickler, Nutzer sowie Betroffene des Systems, um algorithmische Ungleichbehandlung zu mindern.

### **10.1.3 Training zur Bekämpfung kognitiver Verzerrungen**

Da die Entwickler nun Bewusstsein für kognitive Verzerrung aufgebaut haben und diese Verzerrungen die Quelle der meisten Voreingenommenheiten sind, wird im Folgenden *auf Trainingsmaßnahmen eingegangen, mit welchen man diesen entgegenwirken kann*, um somit die Designentscheidungen während des Entwicklungsprozesses zu verbessern (Marques, 2021).

Angefangen mit einem einmaligen Training, welches zur Bekämpfung von kognitiver Voreingenommenheit herangezogen werden kann. Dieses würde in Form eines Videos oder mit einer interaktiven Anwendung erfolgen und wäre eine effektive Strategie, um kognitive Verzerrungen zu bekämpfen (Marques, 2021).

Eine alternative Trainingsstrategie umfasst eine 45-minütige Session, in welcher zuerst das Thema kognitive Verzerrung und Voreingenommenheit näher erläutert und anschließend mit einer Trainingseinheit bekämpft wird. Mit dieser Strategie konnten Voreingenommenheiten für Acht oder mehr Wochen gemindert werden (Marques, 2021).

Generell existieren weitere Möglichkeiten, wie man kognitiver Voreingenommenheit entgegenwirken kann, diese werden hier jedoch nicht näher erläutert, da dies den Rahmen dieser Bachelorarbeit sprengen würde (Marques, 2021).

Nichtsdestotrotz deutlich wird, dass kognitive Verzerrungen schnell und effektiv bekämpft werden können und viele verschiedene Trainingsmaßnahmen existieren, von welchen die für das Entwicklerteam am besten geeigneten gewählt werden könnten. Diese Maßnahmen sollten dann nicht einmalig, sondern in regelmäßigen Abständen erfolgen, da ihre Wirkung nicht für immer, sondern zeitlich begrenzt anhält.

#### 10.1.4 **Expertise**

Es wird angenommen, dass diejenigen KI-Entwickler, die *viel Fachwissen sowie Erfahrung* besitzen, besser darin sind, in ihren Entscheidungen Voreingenommenheiten zu entdecken und diese somit zu mindern, als unerfahrene sowie wenig *Expertenwissen* besitzende KI-Entwickler (Marques, 2021). Auf dieser Grundlage stellte Marques (2021) folgende zwei Hypothesen auf: (a) Personen mit einem höheren selbst eingeschätzten Wissensstand sind besser in der Lage, potenziell verzerrende Variablen in algorithmischen Entscheidungsprozessen zu erkennen und (b) Personen mit einem höheren selbst eingeschätzten Wissensstand berücksichtigen weniger wahrscheinlich potenziell verzerrende Variablen in die algorithmische Entscheidungsfindung ein.

Marques (2021) untersuchte diese Hypothesen und kam zu dem Ergebnis, dass KI-Entwickler mit viel Expertenwissen sowie Erfahrung tatsächlich besser darin sind, voreingenommene Merkmale zu erkennen, sie diese jedoch nicht weniger wahrscheinlich in das KI-System einbeziehen. Hierfür sieht er zwei Gründe. Einerseits können diese KI-Entwickler wie alle anderen Menschen ebenfalls Opfer kognitiver Verzerrungen sein. Andererseits könnte der KI-Entwickler voreingenommene und eventuell sogar diskriminierende Merkmale in das KI-System einbeziehen mit dem Ziel, seine Genauigkeit zu erhöhen. Also würde der Entwickler sich dazu entscheiden, *die Präzision des KI-Modells* auf Kosten von möglicher algorithmischer Diskriminierung zu verbessern.

Hier gilt mein Vorschlag: Solche KI-Entwickler sollten in einem diversen Team arbeiten, so dass mehr Voreingenommenheiten aufgedeckt werden und das Team dem KI-Entwickler die Notwendigkeit der Nichtberücksichtigung mancher möglicherweise diskriminierender Merkmale aufzeigt. Zusätzlich zu einem vielfältigen Team, ist ein Bewusstsein für kognitive- sowie datentechnische Verzerrung und folglich algorithmische Voreingenommenheit anzustreben.

Das Exkludieren solcher Merkmale könnte auf Kosten von algorithmischer Genauigkeit erfolgen. Jedoch könnte dem Entwickler mit Expertise illustriert werden, wie die Ausschließung solcher Merkmale in einem *unvoreingenommenerem und fairerem KI-System* münden kann und dass die mögliche einhergehende Minderung der Genauigkeit somit eventuell in Kauf genommen werden könnte. Dies könnte durch vorhandenes Bewusstsein für algorithmische

Voreingenommenheit verstärkt werden, da der Entwickler mit Expertise sich somit den möglichen verheerenden Auswirkungen bewusst ist und eine Minderung der Genauigkeit somit noch weiter eingehen würde.

Darüber hinaus kann der KI-Entwickler Trainingsstrategien zur weiteren Reduzierung von kognitiver Voreingenommenheit durchführen.

In diesem Kapitel wurde zuerst erläutert, dass ein vielfältigeres Entwicklerteam und die Einbeziehung von Nutzern sowie Betroffenen in die Entwicklung, in einem KI-System münden kann, welches weniger algorithmische Voreingenommenheiten enthält.

Weiterführend wurde illustriert, wie die Vorteile eines solch diversen KI-Entwicklerteams nicht greifen, sofern die Entwickler sich den kognitiven- sowie datentechnischen Verzerrungen und folglich den algorithmischen Voreingenommenheiten nicht bewusst sind. Dieses Bewusstsein bildet die Grundlage zur Bekämpfung kognitiver- sowie algorithmischer Verzerrungen. Dementsprechend sollten sich nicht nur die KI-Entwickler, sondern ebenfalls die Betroffenen sowie insbesondere die Nutzer des Systems sich solch ein Bewusstsein aneignen, da diese Nutzer eventuell als finale Entscheidungsentität fungieren und somit die Handlungsempfehlungen des KI-Systems nicht als neutrale Wahrheit ansehen sollten, sondern diese im Hinblick auf algorithmische Verzerrung, kritisch betrachten sollten.

Diejenigen Entwickler, welche ein Bewusstsein für kognitive Verzerrungen aufgebaut haben, können diese mithilfe von etablierten Trainingsstrategien bekämpfen.

Abschließend wurde erörtert, wie KI-Entwickler mit viel Fachwissen sowie Erfahrung trotzdem diskriminierende Merkmale im KI-System berücksichtigen und wie durch die Maßnahmen eines diverseren Teams, Bewusstsein für Voreingenommenheiten und mit Training zur Bekämpfung kognitiver Verzerrungen dieser Problematik entgegengewirkt werden kann. Dies illustriert, wie die hier vorgestellten Maßnahmen zur Bekämpfung von kognitiven Verzerrungen und darüber hinaus algorithmischen Voreingenommenheiten, am effektivsten wirken, wenn sie *gemeinsam* angewandt werden.

## **10.2 Datentechnische Maßnahmen**

Im Folgenden werden *datentechnische Lösungsmöglichkeiten* erläutert, um den datenbegründeten Ursachen für algorithmische Voreingenommenheit entgegenwirken zu können. Der Datensatz, welcher dem KI-Modell zum Lernen übergeben wird und aus welchem sich die Trainings- sowie Testdaten bilden, wäre diejenige Datenmenge, welche die kognitiven-, gesellschaftlichen sowie branchenspezifischen Voreingenommenheiten widerspiegelt. Weiterführend wäre die Art und Weise, wie das KI-Modell mit den Daten lernt ebenfalls Grundlage für weitere mögliche Ursachen algorithmischer Verzerrung, sodass hierfür in diesem Kapitel ebenfalls entgegenwirkende Maßnahmen vorgestellt werden.

### **10.2.1 Vielfältige sowie ausgewogene Datensätze**

Aktuell erlernen KI-Modelle ihr Verhalten oftmals anhand historischer Daten. Diese historischen Daten können jedoch Voreingenommenheiten enthalten, welche in das Verhalten des AES einfließen könnten. Ein Beispiel hierfür wurde mit dem Amazon-Fallbeispiel angeführt. Die dort an das KI-System übergebene Datengrundlage spiegelte das historische Ungleichgewicht von Frauen und Männern innerhalb der IT-Branche wider. Dementsprechend enthalten diese Daten bestehende kognitive- sowie branchenspezifische Vorurteile. Das lernende KI-Modell berücksichtigt diese Voreingenommenheiten und resultiert in einem finalen Modell, welches Frauen diskriminiert und somit das Ungleichgewicht dieser Branche wiedergibt und weiter verstärkt.

Hervorkommt, dass das *Datenungleichgewicht*, also die Unterrepräsentierung von eingestellten Frauen und die Überrepräsentierung von eingestellten Männern in einem KI-System bzw. AES mündet, welches dieses Ungleichgewicht beibehält, indem es Frauen, lediglich auf Grundlage ihres Geschlechtes, eine geringe Chance zuweist, eingestellt zu werden (Marques, 2021).

Nun ist es jedoch so, dass diese AES verwendet werden, um unter anderem objektivere Entscheidungen zu treffen als Menschen., da Menschen voreingenommen sind.

Nichtsdestotrotz, wenn man dem KI-Modell lediglich vergangene Entscheidungen von Menschen als Grundlage zum Erlernen seines Verhaltens übergibt, dann resultiert dies nicht in einem objektiven sowie zukunftsorientierten KI-System, sondern vielmehr in einem, welches sich die bestehenden Voreingenommenheiten der Menschen aneignet und diese in zukünftige Entscheidungen einbezieht (De Montis, 2019; Machill, 2020; Marques, 2021).

Dementsprechend sind *faire Datensätze* entscheidend. Hiermit ist gemeint, dass *Datensätze möglichst vielfältig und ausgewogen* sein sollten (Fiegele, 2021; Machill, 2020; Marques, 2021).

Auf das Amazon-Beispiel bezogen würde dies konkret bedeuten, dass diejenigen Bewerbungen, welche dem KI-Modell als Trainingsgrundlage zugespielt werden, eine balancierte Menge von Frauen und Männern darstellen sollte.

Also das im Optimalfall die Menge an getätigten und erfolgreichen Bewerbungen gleichmäßig aufgeteilt auf Frauen und Männer ist. Somit würde der Algorithmus nicht zu dem voreingenommenen Entschluss gelangen, dass Frauen für technische Berufe weniger geeignet wären als Männer (Jann & Brunner, 2020).

Dies würde zu einer objektiveren Entscheidungsfindung des Modells beitragen. Sollten nicht genug erfolgreiche Bewerbungen für Frauen vorliegen, so wären *weitere Datenquellen* einzubeziehen oder einige Daten müssten vom Entwicklerteam *zum Zweck des Trainings angepasst* werden (Fiegele, 2021; Marques, 2021).

Das Ziel sollte sein, dass keines der von den Entscheidungen des Systems betroffenen Personengruppen, innerhalb der Trainings- sowie Testdaten, unter- oder überrepräsentiert ist. Auf diese Weise würde sich die Wahrscheinlichkeit dafür verringern, dass sich das AES Voreingenommenheiten aneignet.

Hier wurde aus Einfachheitsgründen auf die Mitberücksichtigung der vielen möglichen alternativen Gendermöglichkeiten verzichtet. In der Praxis wären diese ebenfalls ausreichend im Datensatz zu berücksichtigen. Das generelle Konzept wurde jedoch trotzdem verdeutlicht.

Hierzu ein weiteres Beispiel. Angenommen ein KI-Modell soll prognostizieren, ob jemand eine Straftat begehen wird. Hätte man nun eine geringe Menge an dunkelhäutigen Personen

innerhalb der Trainingsdaten und all diese haben eine Straftat begangen, so könnte das KI-System fälschlicherweise erlernen, dass jede dunkelhäutige Person eine Straftat begehen wird. Somit würde sich das Modell, anhand der Trainingsdaten, eine rassistische Voreingenommenheit aneignen. Dies illustriert die Relevanz eines vielfältigen und ausgewogenen Datensatzes.

Die Erreichung eines solchen Datensatzes ist komplex. Nichtsdestotrotz existieren heutzutage verbesserte Datensätze wie z. B. *Der Deutsche Kredit-Datensatz*. Dieser wurde konzipiert, um die in den historischen Datensätzen vorhandenen Voreingenommenheiten zu korrigieren. Dementsprechend kann ein KI entwickelndes Unternehmen evaluieren, ob es diese Datensätze eventuell verwenden will oder zusätzlich zu ihrem Datensatz als weitere Datenquelle hinzuziehen möchte (Fiegele, 2021; Marques, 2021).

Weiterführend wäre für die Realisierung eines solch vielfältigen und ausgewogenen Datensatzes auch hier ein diverses Entwicklerteam Bewusstsein sowie das Einbeziehen von Nutzern und Betroffenen von hoher Relevanz (De Montis, 2019; Machill, 2020).

Mit diesem Unterkapitel konnte illustriert werden, wie mit einem ausgewogenen und vielfältigen Datensatz einigen der thematisierten datenbegründeten Ursachen für algorithmische Voreingenommenheit, insbesondere Voreingenommenheiten aufgrund von historischen Daten, entgegengewirkt werden kann.

### **10.2.2 Einbeziehung geschützter Merkmale**

Dieser Abschnitt soll hervorheben, wieso die *Einbeziehung von geschützten Merkmalen* in den Datensatz des KI-Systems hinsichtlich der Minderung möglicher entstehender algorithmischer Verzerrungen besser wäre, als dass das System sich andere verfügbarer Daten als Proxy-Variablen also Platzhalter, für diese geschützten Merkmale heranzieht (Mohammad, 2021).

Es wird empfohlen, die Nutzung von Proxy-Variablen zu verhindern, indem die geschützten Merkmale im Datensatz des KI-Systems beibehalten werden. Die Beibehaltung dieser würde die Notwendigkeit einer Proxy-Variable aufheben (Mohammad, 2021).

Diese Empfehlung hat einerseits den Grund, dass algorithmische Voreingenommenheiten schwieriger zu entdecken wären, wenn sie auf Grundlage von Proxys entstanden sind. Andererseits gestaltet sich die Feststellung von Diskriminierungen schwieriger, wenn Proxy-Variablen die Ursache dieser sind (Mohammad, 2021).

Ein Beispiel für eine Ungleichbehandlung aufgrund des Weglassens von geschützten Merkmalen könnte mit dem COMPAS-Fallbeispiel nähergebracht werden. Hier würde das KI-System aufgrund fehlender Berücksichtigung des Geschlechtes Frauen die gleiche Rückfallchance zuweisen wie Männern, obwohl Frauen in der Realität nicht so häufig rückfällig werden wie Männer. Dieses Beispiel verdeutlicht, dass bei Weglassen eines geschützten Merkmals und wenn dieses nicht durch Proxys ersetzt wird, dass dann trotzdem Voreingenommenheiten aufkommen können, da - wie hiermit hervorgehoben - das Einbeziehen von geschützten Attributen nicht nur negative Auswirkungen hat (Kolleck & Orwat, 2020).

Weiterführend würde das Einbeziehen von geschützten Merkmalen darin münden, dass diese präventiv verwendet werden können, um algorithmische Diskriminierung zu mindern (Mohammad, 2021).

Außerdem, wie im Unterkapitel „Proxy-Variablen“ bereits erörtert, wäre wiederholt hervorzuheben, dass die Entdeckung einer Verzerrung verursachenden Proxy-Variable und die darauf folgende Eliminierung dieser, nicht dafür sorgen würde, dass das geschützte Merkmal, für welches die Proxy-Variable als Platzhalter verwendet wurde, ebenfalls aus der Entscheidungsfindung des Systems entfernt wird. Dies hat den Grund, dass Proxy-Variablen von anderen Proxys ersetzt werden, sodass die Nutzung des eliminierten Platzhalters fortbestehen bleibt und somit eventuell auch die mit dieser Nutzung einhergehenden Verzerrungen.

Abschließend lässt sich also sagen, dass das Entwicklerteam geschützte Merkmale verwenden sollte, da sich deren Nutzung gegenüber der Nutzung von Proxys vorteilhafter darstellt und die Einbeziehung dieser geschützten Merkmale die Verwendung von Proxy-Variablen minimieren oder eventuell sogar verhindern würde.

In diesem Kapitel wurden mögliche datentechnische Maßnahmen, die ein KI-Entwicklerteam zur Entwicklung eines möglichst objektiven KI-Systems bzw. AES vornehmen kann, vorgestellt.

Es existieren noch weitere datentechnische Methoden, die jedoch erst im übernächsten Kapitel vorgestellt werden können, da im nächsten Kapitel erst die Metriken erläutert werden, welche von diesen weiteren Methoden verwendet werden.

Wichtig an dieser Stelle hervorzuheben, ist dass die bereits thematisierten Feedback-Schleifen bei einem System mit voreingenommenem und diskriminierendem Verhalten dafür sorgen würden, dass sich dieses negative Verhalten vertieft sowie festsetzt.

Nichtsdestotrotz würden diese Feedback-Schleifen das Verhalten eines objektiven KI-Systems, welches möglichst unvoreingenommen und diskriminierungsfrei entscheidet, verstärken. Dies hebt die Relevanz dieser datentechnischen Maßnahmen hervor, da diese das Verhalten des KI-Modells initial in eine möglichst unvoreingenommene und diskriminierungsfreie Richtung lenken würden und dieses KI-System bzw. AES mithilfe seiner Feedback-Schleifen dieses Verhalten beibehalten und sogar vertiefen würde.

### **10.3 Statistische Metriken zur Messung der Genauigkeit & Fairness**

Damit algorithmische Voreingenommenheit mindernde Maßnahmen überhaupt durchgeführt werden können, benötigt es einen Verdacht darauf, dass das KI-System gleich zu behandelnde Personen ungleich behandelt und das System somit ein voreingenommenes sowie eventuell diskriminierendes Verhalten aufweist (Quelle: (Machill, 2020)).

Dementsprechend existieren *statistische Metriken*, um Hinweise auf die *Genauigkeit und Fairness* eines KI-Systems zu erlangen (Quelle: (Machill, 2020)).

Bevor diese Metriken genauer thematisiert werden, wäre es zum Verständnis notwendig, die Begriffe *geschützte Gruppe* und *ungeschützte Gruppe* näher zu erläutern (Verma & Rubin, 2018).

(Verma & Rubin, 2018) und Rubin haben diese Begriffe mit den Geschlechtern in Verbindung gebracht. In ihrem wissenschaftlichen Artikel wurden Männer in die ungeschützte Gruppe und Frauen in die geschützte Gruppe eingeordnet (Machill, 2020; Verma & Rubin, 2018).



Dies lässt sich anhand der vorherrschenden und in dieser Arbeit an einigen Stellen ebenfalls deutlich gemachten Ungleichbehandlung von Mann und Frau, erklären (Pawlik, 2022).

Dementsprechend kann man sich unter geschützten Gruppen diejenigen Personen vorstellen, die eher benachteiligt werden und somit Schutz vor Diskriminierung erhalten und beziehen sollten im Gegensatz zu denjenigen, die eher Privilegien genießen und dementsprechend eventuell keinen besonderen Schutz vor Diskriminierung beziehen müssen, und somit in die ungeschützte Gruppe eingeordnet werden.

### **10.3.1 Confusion-Matrix & Allgemeine statistische Metriken**

Nun zu den Metriken: Zuerst werden die *Kern-Metriken* der *Confusion Matrix* vorgestellt. Diese Matrix wird dazu verwendet, um festzustellen wie gut ein Klassifikationsmodell ist. Hierfür kann mithilfe der Confusion Matrix die Genauigkeit des Modells ermittelt werden (Machill, 2020; Narkhede, 2018).

Diese Matrix wäre demzufolge im Kontext von KI weit verbreitet. Sie wäre in Abbildung 14 dargestellt und enthält Kern-Metriken, auf welche im Folgenden näher eingegangen wird (Machill, 2020; Verma & Rubin, 2018).

		Tatsächliches Ergebnis	
		Positiv	Negativ
Vorhergesagtes Ergebnis	Positiv	TP	FP
	Negativ	FN	TN

Abbildung 14: Confusion Matrix (eigene Darstellung) (Narkhede, 2018)

Die Matrix fokussiert sich auf die positiven und negativen Vorhersagen und Ergebnisse des Modells. Ein positives Ergebnis wäre hierbei beispielsweise, dass eine Person einen Kredit erhalten sollte. Während ein negatives Ergebnis beispielsweise wäre, dass eine Person keinen Kredit erhalten sollte. Gleiches gilt für die Vorhersagen des Modells (Verma & Rubin, 2018).

Um die Güte bzw. Genauigkeit des Modells letztlich ermitteln zu können, werden die vorhergesagten positiven/negativen Ergebnisse und die tatsächlich positiven/negativen Ergebnisse in Beziehung gebracht. Folglich kommen vier Kern-Metriken zum Vorschein, welche wären, *True Positive (TP)*, *False Positive (FP)*, *False Negative (FN)* und *True Negative (TN)* (Verma & Rubin, 2018).

Man spricht von TP, wenn eine getroffene positive Vorhersage des Modells tatsächlich positiv ist. Also in unserem Beispiel, wenn das Modell prognostiziert, dass eine Person einen Kredit erhalten sollte und diese Person tatsächlich einen Kredit erhalten hat (Verma & Rubin, 2018).

Man spricht von FP, wenn eine getroffene positive Vorhersage des Modells tatsächlich negativ ist. Also in unserem Beispiel, wenn das Modell prognostiziert, dass eine Person einen Kredit erhalten sollte und diese Person aber tatsächlich keinen Kredit erhalten hat (Verma & Rubin, 2018).

Man spricht von FN, wenn eine getroffene negative Vorhersage des Modells tatsächlich positiv ist. Also in unserem Beispiel, wenn das Modell prognostiziert, dass eine Person keinen Kredit erhalten sollte und diese Person aber tatsächlich einen Kredit erhalten hat (Verma & Rubin, 2018).

Man spricht von TN, wenn eine getroffene negative Vorhersage des Modells tatsächlich negativ ist. Also in unserem Beispiel, wenn das Modell prognostiziert, dass eine Person keinen Kredit erhalten sollte und diese Person tatsächlich keinen Kredit erhalten hat (Verma & Rubin, 2018).

Diese Kern-Metriken werden als Fundament für die Ermittlung von *allgemeinen statistischen Metriken* herangezogen, welche im Folgenden in Tabelle 1 vorgestellt werden (Machill, 2020).

Metrik	Formel	Erläuterung
<i>Precision</i>	$TP / (TP + FP)$	Misst, wie wahrscheinlich es ist, dass eine positive Vorhersage des Modells tatsächlich positiv ist.
<i>False discovery rate (FDR)</i>	$FP / (TP + FP)$	Misst, wie wahrscheinlich es ist, dass eine positive Vorhersage des Modells tatsächlich negativ ist.
<i>False omission rate (FOR)</i>	$FN / (TN + FN)$	Misst, wie wahrscheinlich es ist, dass eine negative

		Vorhersage des Modells tatsächlich positiv ist.
<i>Negative predictive value (NPV)</i>	$TN / (TN + FN)$	Misst, wie wahrscheinlich es ist, dass eine negative Vorhersage des Modells tatsächlich negativ ist.
<i>Recall</i>	$TP / (TP + FN)$	Ermittelt, wie wahrscheinlich ein tatsächlich positives Ergebnis auch als solches vom Modell vorhergesagt wird.
<i>False positive rate (FPR)</i>	$FP / (FP + TN)$	Ermittelt, wie wahrscheinlich ein tatsächlich negatives Ergebnis vom Modell als positives vorhergesagt wird.
<i>False negative rate (FNR)</i>	$FN / (TP + FN)$	Ermittelt, wie wahrscheinlich ein tatsächlich positives Ergebnis vom Modell als negatives vorhergesagt wird.
<i>True negative rate (TNR)</i>	$TN / (FP + TN)$	Ermittelt, wie wahrscheinlich ein tatsächlich negatives Ergebnis auch als solches vom Modell vorhergesagt wird.
<i>Accuracy</i>	$(TP + TN) / (TP + TN + FP + FN)$	Ermittelt einen prozentualen Wert, welcher wiedergibt, wie viele Vorhersagen aller

		Vorhersagen des Modells richtig waren.
--	--	--

Tabelle 1: Allgemeine statistische Metriken (Narkhede, 2018; Oppermann, 2021; Verma & Rubin, 2018)

Mithilfe dieser vorgestellten Metriken kann die *Genauigkeit bzw. Korrektheit* des prognostizierenden KI-Systems festgestellt werden.

### 10.3.2 Fairness-Metriken

Weiterführend werden im Folgenden erweiterte Metriken illustriert, welche speziell zur Ermittlung der Fairness beitragen.

Es existieren zahlreiche solcher *Fairness-Metriken*, welche auf den allgemeinen Metriken beruhen. Diese Arbeit fokussiert sich jedoch auf diejenigen, welche in der Forschung die meiste Relevanz genießen (s. Abb. 15) (Machill, 2020; Verma & Rubin, 2018).

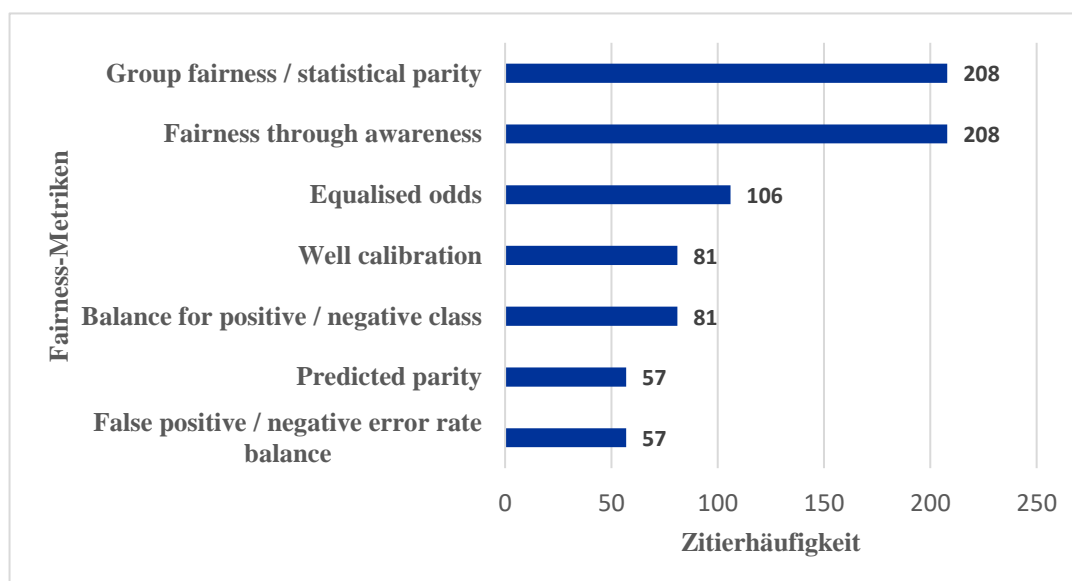


Abbildung 15: Meistzitierten Fairness-Metriken (eigene Darstellung) (Machill, 2020; Verma & Rubin, 2018)

Diese werden im Folgenden vorgestellt.

Angefangen mit der *Group fairness* oder auch *statistical parity*. Diese Metrik wird erreicht, wenn Personen der ungeschützten- sowie geschützten Gruppen die gleiche Chance darauf haben, vom Modell in die positive Klasse eingeordnet zu werden (Machill, 2020; Verma & Rubin, 2018).

Auf das bereits vorgestellte Beispiel bezogen würde dies bedeuten, dass das Modell für Frauen sowie für Männer mit gleicher Wahrscheinlichkeit vorhersagen würde, einen Kredit zu erhalten (Machill, 2020; Verma & Rubin, 2018).

Diese Metrik weist jedoch Mängel auf. Sie kann keine Erhebungen zu Kombinationen von Gruppen durchführen, sondern ist nur in der Lage, Gruppen je einzeln zu berücksichtigen (Machill, 2020; Verma & Rubin, 2018).

Auf das vorgestellte Beispiel bezogen würde dies bedeuten, dass mit dieser Metrik keine Überprüfung auf Fairness bezüglich beispielsweise *dunkelhäutiger Frauen* durchgeführt werden kann. Sondern entweder nur nach *Frauen* oder *dunkelhäutigen Personen* überprüft werden kann (Machill, 2020; Verma & Rubin, 2018).

Dies würde verhindern, dass Diskriminierungen, welche nur auf eine Kombination von Gruppen ausgeübt werden, entdeckt werden würden (Machill, 2020).

Ein weiterer Mangel wäre die Nichtberücksichtigung von anderen Merkmalen (Machill, 2020).

Auf das bereits vorgestellte Beispiel bezogen würde dies bedeuten, dass Frauen zwar die gleiche Chance haben können, in die positive Klasse eingeordnet zu werden, sie jedoch eventuell beispielsweise mehr Arbeitsjahre als Männer erlebt haben müssen, um die gleichen Chancen wie diese zu erlangen (Machill, 2020).

Aufgrund des Weglassens von Merkmalen kann diese Metrik folglich bestehende Ungleichbehandlungen, welche nur mithilfe von zusätzlichen Variablen erkenntlich werden würden, nicht ermitteln (Machill, 2020).

Die nächste Fairness-Metrik namens *Fairness through awareness* würde, anders als die Group fairness, sich auf Individuen fokussieren. Hiermit könnten die in der statistical parity angesprochenen Mängel, bekämpft werden (Machill, 2020; Verma & Rubin, 2018).

Fairness through awareness besagt, dass ähnliche Individuen auch eine ähnliche Einordnung erfahren sollten. Hierfür würde die Gleichheit/Ungleichheit von Individuen mittels einer Distanzmessgröße festgestellt werden und, um ein faires Modell aufweisen zu können, sollte die Differenz der Verteilung der prognostizierten Ergebnisse geringer oder maximal genauso groß sein wie die Differenz der Individuen selbst (Machill, 2020; Verma & Rubin, 2018).

Auf das Beispiel bezogen würde dies bedeuten, dass zehn Personen, die exakt identische Merkmale aufweisen, sich aber im Geschlecht unterscheiden, trotzdem alle die gleiche Prognose zugeteilt bekommen sollten (Machill, 2020; Verma & Rubin, 2018).

Sollte die Prognose in diesem Falle z. B. für Frauen schlechter sein, so wird eine geschlechter-spezifische Diskriminierung deutlich (Machill, 2020; Verma & Rubin, 2018).

Diese beiden Metriken stellen mit Abstand die zwei meistverwendeten Fairness-Metriken dar (s. Abb. 15).

Die nach Group Fairness und Fairness through awareness am dritthäufigsten verwendete Fairness-Metrik wäre *Equalized odds* (s. Abb. 15).

Diese Metrik soll sicherstellen, dass das Modell, unabhängig davon, welche Gruppe es vorliegen hat, gleich gute Prognosen abliefert und demzufolge über alle Gruppen hinweg eine gleichbleibende Genauigkeit der Vorhersagen beibehält (*Common fairness metrics*, o. D.; Machill, 2020).

Hierfür prüft diese Metrik, ob ungeschützte- sowie geschützte Gruppen die gleiche FPR sowie den gleichen Recall aufweisen (s. Tabelle 1) (*Common fairness metrics*, o. D.; Machill, 2020).

Diese Metrik würde ebenfalls insbesondere gut diejenigen Modelle erkennen, welche dazu neigen, nur bei den ungeschützten Gruppen, also der Mehrheit, gut zu performen (*Common fairness metrics*, o. D.; Machill, 2020).

Aus diesen Gründen eignet sich Equalized odds exzellent als Fairness-Metrik (*Common fairness metrics*, o. D.; Machill, 2020).

Die nächste Fairness-Metrik namens *Well-calibration* wird erreicht, wenn das Modell die Individuen der geschützten Gruppen mit exakt gleicher Wahrscheinlichkeit in die positive Klasse eingeordnet wie diejenigen Individuen der ungeschützten Gruppen (Verma & Rubin, 2018).

Auf das Beispiel bezogen würde dies bedeuten, dass Frauen sowie Männer die exakt gleiche Wahrscheinlichkeit dafür haben, in die positive Klasse eingeordnet zu werden (Verma & Rubin, 2018).

Weiterführend existiert die Metrik *Balance for positive class*. Diese Metrik wird erreicht, wenn diejenigen Individuen aus den ungeschützten- sowie geschützten Gruppen, welche tatsächlich zur positiven Klasse gehören, mit gleicher durchschnittlicher Wahrscheinlichkeit dieser Klasse zugeordnet werden (Machill, 2020; Verma & Rubin, 2018).

Gegenteilig wird die Metrik *Balance for negative class* erreicht, wenn diejenigen Individuen aus den ungeschützten- sowie geschützten Gruppen, welche tatsächlich zur negativen Klasse gehören, mit gleicher durchschnittlicher Wahrscheinlichkeit dieser Klasse zugeordnet werden (Machill, 2020; Verma & Rubin, 2018).

Weiter mit der Fairness-Metrik *Predictive parity*. Diese ist erfüllt, wenn die positiven Vorhersagen des Modells für ungeschützte- sowie geschützte Gruppen gleich präzise sind (Machill, 2020; Verma & Rubin, 2018).

Auf unser Beispiel bezogen, würde dies bedeuten, dass das Modell für Frauen sowie Männer die gleiche Precision aufweist (s. Tabelle 1) (Machill, 2020; Verma & Rubin, 2018).

Die nächste Metrik wäre *False positive error rate balance*. Diese Metrik wird erreicht, wenn die False positive rate (FPR) des Modells für ungeschützte- sowie geschützte Gruppen gleich hoch ist (s. Tabelle 1) (Machill, 2020; Verma & Rubin, 2018).

Auf unser Beispiel bezogen würde dies bedeuten, dass die Wahrscheinlichkeit des Modells, ein tatsächlich negatives Ergebnis (erhält keinen Kredit) fälschlicherweise als positives (erhält Kredit) vorherzusagen, für Frauen sowie Männer gleich hoch sein sollte (Machill, 2020; Verma & Rubin, 2018).



Konträr existiert die Metrik *False negative error rate balance*. Diese Metrik wird erreicht, wenn die False negative rate (FNR) des Modells für ungeschützte- sowie geschützte Gruppen gleich hoch ist (s. Tabelle 1) (Machill, 2020; Verma & Rubin, 2018).

Auf unser Beispiel bezogen würde dies bedeuten, dass die Wahrscheinlichkeit des Modells, ein tatsächlich positives Ergebnis (erhält Kredit) fälschlicherweise als negatives (erhält keinen Kredit) vorherzusagen, für Frauen sowie Männer gleich hoch sein sollte (Machill, 2020; Verma & Rubin, 2018).

Hiermit wurden die meistverwendeten Fairness-Metriken vorgestellt. Mit diesen kann die Gleich- oder Ungleichbehandlung der Betroffenen des KI-Systems ermittelt werden.

In diesem Kapitel wurde erläutert, wie auf Grundlage der Confusion Matrix und ihren Kern-Metriken allgemeine statistische Metriken erzeugt werden können.

Weiterführend wurde aufgezeigt, wie mithilfe dieser allgemeinen Metriken einerseits die Korrektheit des KI-Modells erhoben werden kann, und andererseits, wie aus diesen Fairness-Metriken, welche speziell zur Ermittlung der Fairness eines KI-Modells entworfen wurden, gebildet werden können.

Auf Grundlage dieser Fairness-Metriken können im nächsten Kapitel weitere datentechnische Maßnahmen vorgestellt werden, welche ebenfalls das Ziel verfolgen, algorithmische Voreingenommenheit zu mindern.

## **10.4 Weitere datentechnische Maßnahmen**

Im Folgenden werden Techniken vorgestellt, welche auf Grundlage des Datensatzes des zu entwickelnden KI-Systems vom Entwicklerteam unter anderem zur Minderung von algorithmischer Voreingenommenheit durchgeführt werden können (Machill, 2020).

Diese Techniken wären *Pre-Processing*, *In-Processing* und *Post-Processing* (Machill, 2020).

### 10.4.1 Pre-Processing

Angefangen mit *Pre-Processing* oder auch der *Datenvorverarbeitung* (Lawton, 2023).

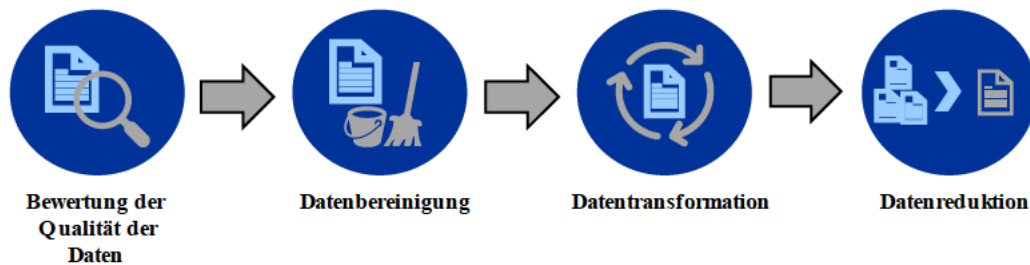


Abbildung 16: Schritte der Datenvorverarbeitung (eigene Darstellung) (Lawton, 2023; Mesevage, 2021)

Der erste Schritt der Datenvorverarbeitung ist *die Bewertung der Qualität der Daten* (s. Abb. 16). Die Entwickler sollen die vorhandenen Daten analysieren und dabei ihre Güte sowie Relevanz für das zu trainierende KI-System evaluieren. Außerdem sind auftretende Datenanomalien zu beachten (Mesevage, 2021).

Diese Anomalien umfassen:

- *nicht übereinstimmende Datentypen*, welche vereinheitlicht werden müssen (Mesevage, 2021).
- *Unterschiedliche Datenwerte*, welche sich syntaktisch voneinander unterscheiden, aber semantisch das Gleiche meinen (Mesevage, 2021).
- *Ausreißer*, welche einzelne Daten darstellen, die das Bild des Datensatzes erheblich verzerren (Mesevage, 2021).
- *Fehlende Daten*, welche auf eine eventuell unvollständige Datenquelle hinweisen (Mesevage, 2021).
- *Verrauschte Daten*, welche neben relevanten Informationen von Datensätzen existieren und weniger relevante, gar bedeutungslose Informationen darstellen, die das Bild des Datensatzes verzerren (Mesevage, 2021; *What is Noise in Data Mining*, o. D.).

Einige dieser Datenanomalien können in dem nächsten Schritt, der *Datenbereinigung* (s. Abb. 16), behoben werden. In diesem Schritt werden insbesondere korrigierende Maßnahmen für fehlende- sowie verrauschte Daten durchgeführt (Lawton, 2023; Mesevage, 2021).

Die zwei meistverwendeten Maßnahmen, um mit fehlenden Daten umzugehen, wäre entweder diese manuell nachzuführen, oder diejenigen Daten-Tupel, welche unvollständig sind, komplett zu entfernen. Das manuelle Nachführen kann jedoch bei zu großen Datenmengen zu aufwändig werden und das Entfernen von ganzen Daten-Tupeln wäre bei zu kleinen Datenmengen nicht zu empfehlen (Mesevage, 2021).

Um mit verrauschten Daten umzugehen, können drei Maßnahmen herangezogen werden (Mesevage, 2021).

Bei der ersten Initiative namens *Binning* würden die Daten des großen vielfältigen Datensatzes in kleinere homogenere Datensätze aufgeteilt werden. Somit werden die bedeutungslosen Daten von den relevanten getrennt (Mesevage, 2021).

Die zweite Maßnahme wäre die *Regression*. Hier wird entschieden, welche Variablen letzten Endes tatsächlich verwendet werden. Dies würde bedeutungslose Merkmale sowie Daten außen vor lassen (Mesevage, 2021).

Die dritte und letzte Möglichkeit wäre *Clustering*. Hier würden ähnliche Daten gruppiert werden. Die Ähnlichkeit wird auf Grundlage der vorhandenen Daten ermittelt, weswegen sich diese Maßnahme gut für die unüberwachende Lernmethode von KI-Systemen eignet (Der Clustering Guide: Definition, Methoden und Beispiele, o. D.; Mesevage, 2021).

Hiermit wurden die möglichen Maßnahmen, welche innerhalb der Datenbereinigungsphase durchgeführt werden können, deutlich gemacht.

Der nächste Schritt wäre die *Datentransformation* (s. Abb. 16). Hier werden die Daten optimal für das Training des KI-Systems angepasst (Mesevage, 2021).

Hier kommen ebenfalls unterschiedliche Maßnahmen zum Einsatz (Mesevage, 2021).

Angefangen mit der *Normalisierung*. Hier werden die Daten in vergleichbarere Bereiche skaliert (Mesevage, 2021).

Weiter mit der *Auswahl der Merkmale* (Mesevage, 2021). Diese Maßnahme wurde im Unterkapitel „Wahl und Gewichtung der Merkmale“ bereits näher erläutert.

Darüber hinaus existiert die *Diskreditierung*. Hier würden Daten in kleinere Intervalle geordnet werden (Mesevage, 2021).

Weiterführend existiert die *Erstellung einer Konzepthierarchie*. Hier können der bereits vorhandenen Hierarchie weitere Stufen angefügt werden. Somit könnten große allgemein gehaltene Gruppen in feingranularere Untergruppen zerteilt werden, welche die Diskriminierung die solch eine Untergruppe möglicherweise erfährt, hervorstechen lassen würde (Machill, 2020; Mesevage, 2021).

Zusätzlich existiert an dieser Stelle eine algorithmische Voreingenommenheit mindernde Maßnahme namens *Massaging*. Hier würde man diejenige Kategorie, in der Klassenbezeichnung, die einem Datensatz aktuell zugeordnet ist, umändern. Beispielsweise würde somit einer Person anstatt „erhält Kredit“ die Kategorie „erhält keinen Kredit“ zugeordnet werden (Kamiran & Calders, 2011; Machill, 2020).

Hier spielt die Auswahl derjenigen Daten, welche umgeändert werden sollen, eine große Rolle. Die Auswahl beruht auf einem *Ranking-System*, sodass diejenigen Daten, welche am weitesten oben im Ranking stehen, am ehesten geändert werden. Hierbei stehen diejenigen Daten weit oben im Ranking, die am ehesten zur anderen Klasse zugeordnet werden. Die Änderung der Kategorie bei diesen Daten hätte nur eine leichte Beeinträchtigung der Genauigkeit der Vorhersagen des KI-Systems zur Folge. Nichtsdestotrotz ist dies eine der effektivsten Methodiken zur Erreichung eines fairen KI-Systems bzw. AES (Kamiran & Calders, 2011; Machill, 2020).

Abschließend, existiert mit *Reweighting* eine Alternative zum Massaging-Ansatz. Hier bekommen Daten eine Gewichtung zugeteilt. Folglich werden diejenigen Daten, die einen bevorzugen oder diskriminierenden Charakter aufweisen, nicht umgeändert, sondern stattdessen bekommen diesen Daten eine geringere Gewichtung zugewiesen, während konträr neutralere Daten eine stärkere Gewichtung zugeteilt bekommen. Je nach Gewichtung bezieht das KI-System die jeweiligen Daten mehr oder weniger in das Erlernen seines Verhaltens ein (Kamiran & Calders, 2011; Machill, 2020).

Auf diese Weise würde sich das KI-System eher auf diejenigen Daten fokussieren, die ihm ein neutrales und faires Verhalten erlernen lassen würden, anstatt sich auf diejenigen zu fokussieren, die ihm ein diskriminierendes Verhalten aneignen lassen (Kamiran & Calders, 2011; Machill, 2020).

Diese Alternative liefert anders als der Massaging-Ansatz eine Möglichkeit der Minderung von algorithmischer Voreingenommenheit, ohne die initialen Ausprägungen der Daten zu verändern, sondern lediglich mit Hinzunahme eines neuen Merkmals, welches jedoch ausschließlich zur Bekämpfung von algorithmischer Voreingenommenheit hinzugezogen wurde (Machill, 2020).

Nichtsdestotrotz haben diese zuletzt vorgestellten Möglichkeiten beide ein unvoreingenommeneres und korrekter arbeitendes KI-System zum Ergebnis (Kamiran & Calders, 2011; Machill, 2020).

Hiermit wurden die in der Datentransformation durchführbaren möglichen Maßnahmen, insbesondere diejenigen, welche algorithmischer Voreingenommenheit entgegenwirken, vorgestellt.

Der nächste und letzte Schritt der Datenvorverarbeitung wäre die *Data reduction* oder auch *Datenreduktion* (s. Abb. 16) (Mesevage, 2021).

Das Ziel dieses Schrittes wäre, die für das Training des KI-Systems wirklich relevanten Daten herauszufiltern, um somit die Analyse der Daten noch einfacher sowie präziser gestalten zu können (Mesevage, 2021).

Eine Maßnahme zur Erfüllung dieses Ziels könnte man mit der *Auswahl der Attribute* erreichen. Dies sollte nicht mit der Auswahl der Merkmale verwechselt werden, denn mit der Auswahl der Attribute ist hier die Handlungsmöglichkeit gemeint, verschiedene Merkmale zu einem Merkmal zusammenzufassen (Mesevage, 2021).

Eine weitere Möglichkeit der Datenreduktion wäre mit der *Dimensionality reduction* oder auch *Verringerung der Dimensionalität* gegeben. Hier könnte man beispielsweise die *Nächste-Nachbarn-Klassifikation* verwenden, um ähnliche Daten zu identifizieren sowie

zusammenzuführen und somit Tausende Merkmale mit nur wenigen transformierten darzustellen (Mesevage, 2021; Thakur, 2021).

Abschließend würde eine ebenfalls sehr effektive Alternative zur Massaging-Maßnahme, mit der *Preferential Sampling*-Methodik vorliegen. Diese Methodik würde, im Kontrast zum Massaging-Ansatz, diejenigen Daten, die im Ranking weit oben stehen, nicht in eine andere Kategorie einordnen, sondern diese entweder duplizieren oder ganz entfernen. Auf diese Weise kann algorithmischer Voreingenommenheit entgegengewirkt werden und die Größe des Datenbestands kann nach Wunsch entsprechend angepasst werden (Machill, 2020).

Hiermit wurden die in der Datenreduktion durchführbaren möglichen Maßnahmen, insbesondere diejenigen, welche algorithmischer Voreingenommenheit entgegenwirken, vorgestellt.

Die Technik des Pre-Processing wurde somit erläutert. Verdeutlicht wurden Maßnahmen, die der Minderung von algorithmischer Verzerrung dienen und durch das Hinzufügen, Löschen, Duplizieren sowie Modifizieren der Daten erreicht werden können.

#### 10.4.2 In-Processing

Die nächste Technik namens *In-Processing* hätte ebenfalls zum Ziel, ein korrekt sowie fair arbeitendes KI-Modell zu erreichen, jedoch anders als beim Pre-Processing-Ansatz, würden beim In-Processing dabei keine Daten verändert werden, sondern vielmehr zielt diese Technik darauf ab, den Lernprozess des KI-Systems anzupassen (Machill, 2020).

Aufgrund der Tatsache, dass zu den einzelnen Schritten dieser Technik keine zitierfähigen Quellen identifiziert werden konnten, wird im Folgenden lediglich eine mögliche algorithmische Voreingenommenheit mindernde Maßnahme vorgestellt, die beim In-Processing Verwendung findet.

Diese Maßnahme lautet *Adversarial Debiasing*. Hier wird eine Lernmethode beschrieben, bei welcher ein sogenannter *predictor*, welcher als unser KI-Modell verstanden werden kann, auf Grundlage einer Eingabe, eine Ausgabe ermitteln soll. Die Eingaben wären die dem Modell

übergebenen Daten und die Ausgaben wären die resultierenden Vorhersagen (Machill, 2020; Zhang, Lemoine & Mitchell, 2018)

Der predictor verwendet ebenfalls Gewichtungen für seine Variablen und passt diese an, um die Differenz zwischen prognostiziertem- und tatsächlichem Wert zu minimieren und somit folglich möglichst akkurate Ergebnisse vorherzusagen (Zhang et al., 2018).

Nun würde jedoch zusätzlich zum predictor eine weitere Komponente namens *adversary* Anwendung finden. Diese Komponente wäre ein separates KI-Modell, welches die Fairness des predictor-Modells prüfen soll (Zhang et al., 2018).

Hierfür werden Fairness-Metriken wie die group fairness oder equalized odds hinzugezogen (Machill, 2020; Zhang et al., 2018).

Sollte der adversary also zum Ziel haben zu prüfen, ob der predictor die Metrik group fairness erfüllt, so würde dieser die Vorhersagen des predictors als Eingabe verwenden, um zugrundeliegende geschützte Merkmale auf Grundlage dieser Eingabe zu prognostizieren (Zhang et al., 2018).

Im Falle, dass der adversary auf Grundlage der Vorhersagen des predictors korrekt darauf verweisen kann, dass beispielsweise das Merkmal Nationalität bei der Prognosefindung des predictors verwendet worden ist, so wäre dies ein Hinweis darauf, dass die Prognose bzw. Prognosefindung des predictors diskriminierendes Verhalten aufweisen kann (Zhang et al., 2018).

Folglich hätte der predictor zum Ziel, den adversary daran zu hindern, solche Vorhersagen treffen zu können (Zhang et al., 2018).

Demensprechend würde der predictor seine Entscheidungsfindung so anpassen, dass stets eine möglichst genaue Vorhersage getroffen wird, aber die Entscheidungsfindung gleichzeitig kein voreingenommenes oder diskriminierendes Verhalten aufweist, sodass der adversary nicht in der Lage sein kann, geschützte Merkmale anhand der Prognosen des predictors zu ermitteln (Zhang et al., 2018).

Dies erreicht der predictor, indem er sein Ziel, nämlich die Differenz zwischen prognostiziertem- und tatsächlichem Wert zu minimieren, erweitert und in die Gewichtung seiner

Variablen zusätzlich noch den adversary mitberücksichtigt. Folglich würde der predictor versuchen, den adversary daran zu hindern, seine Differenz zwischen prognostiziertem- und tatsächlichem Wert zu minimieren, sodass dieser adversary, im Optimalfall, keine geschützten Merkmale prognostizieren kann (Zhang et al., 2018).

Konkret würde der predictor diese Ziele durch Anpassung seiner Entscheidungsfindung, bzw. Anpassung seiner Gewichtungen, erreichen. Hierfür könnte der predictor seine Gewichtungen dahingehend anpassen, dass die vom adversary ermittelte Differenz zwischen prognostiziertem- und tatsächlichem Wert möglichst hoch ist (Zhang et al., 2018).

Die Hinzunahme des adversary würde den predictor dazu zwingen, seine Vorhersagen weniger auf Grundlage von geschützten Merkmalen oder Proxys zu fällen, sondern diese vielmehr auf Basis von objektiven Variablen zu treffen (Zhang et al., 2018).

Auf diese Art und Weise würde die Lernmethodik des Adversarial Debiasing die Vorhersagegenauigkeit des predictors zwar möglicherweise beeinträchtigen, aber dafür hätte man somit ein unvoreingenommeneres und faireres KI-Modell zum Ergebnis, welches Fairness-Metriken wie group fairness oder equalized odds erfüllen würde (Machill, 2020; Zhang et al., 2018).

Die Technik des In-Processing bzw. die Lernmethodik des Adversarial Debiasing wurde somit ausgiebig erläutert. Verdeutlicht wurde, wie mit dieser Maßnahme algorithmische Verzerrung mitigiert werden kann (Zhang et al., 2018).

### **10.4.3 Post-Processing**

Abschließend wird die *Post-Processing*-Technik vorgestellt. Diese hätte ebenfalls zum Ziel, ein korrekt sowie fair arbeitendes KI-Modell zu erreichen, jedoch anders als beim Pre-Processing oder In-Processing werden beim Post-Processing keine Daten oder der Lernprozess des KI-Modells verändert. Dies hat den Grund, dass sich der Post-Processing-Ansatz auf KI-Systeme fokussiert, die als Black-Box verstanden werden können, wodurch hier keine Änderungen wie beim Pre-Processing oder In-Processing beschrieben, realisiert werden können (Machill, 2020).



Diese Post-Processing-Algorithmen oder *Nachbearbeitungsalgorithmen* zielen vielmehr darauf ab, die Korrektheit sowie Fairness des KI-Systems zu erhöhen, nachdem dieses initial fertig trainiert worden ist (Lohia et al., 2018; Machill, 2020).

Ähnlich zum In-Processing-Unterkapitel liegt beim Post-Processing-Unterkapitel der Fokus ebenfalls auf einer ausgewählten möglichen Methodik zur Minderung von algorithmischer Voreingenommenheit.

Der von Lohia et al. entworfene Fairness-Nachverarbeitungsalgorithmus hat zum Ziel, individuelle- sowie group fairness zu erfüllen. Unter individueller Fairness kann man sich die *Fairness through awareness*-Metrik vorstellen (Lohia et al., 2018).

Dieser Algorithmus funktioniert wie folgt: Zuerst werden jegliche Daten der Minderheiten mithilfe eines Vorurteilsdetektors auf individuelle Vorurteile geprüft. Sollte individuelle Voreingenommenheit vorliegen, so würde diesen Daten das Ergebnis zugeteilt werden, welches sie gehabt hätten, wenn sie keine Minderheit wären, sondern zur Mehrheit gehört hätten. Dies hätte wahrscheinlich, ähnlich zum Massaging-Ansatz, eine Änderung der Kategorie dieser betroffenen Daten zur Folge (Lohia et al., 2018).

Auf diese Weise würde der von Lohia et al., 2018 erstellte Fairness-Nachverarbeitungsalgorithmus versuchen sicherzustellen, dass Minderheiten nicht aufgrund von individuellen Vorurteilen eine benachteiligende Vorhersage erleiden, sondern diese vielmehr die Chance darauf haben eine faire Vorhersage zu erfahren (Lohia et al., 2018).

Mit diesem Unterkapitel wurde erläutert, wie mit der Technik des Post-Processing bzw. dem von Lohia et al. erstellten Post-Processing-Algorithmus algorithmischer Voreingenommenheit entgegengewirkt werden kann und wie somit die Fairness-Metriken *Fairness through awareness* und *group fairness* erfüllt werden (Lohia et al., 2018).

All die in Kapitel „Statistische Metriken zur Messung der Genauigkeit & Fairness“ vorgestellten Metriken sowie die hier erläuterten weiteren datentechnischen Maßnahmen können während der KI-Entwicklung verwendet werden, um ein objektiveres und korrekter arbeitendes KI-System bzw. AES hervorzubringen (Lohia et al., 2018).

Nichtsdestotrotz wären diese vorgestellten Metriken sowie Techniken generell nur für die überwachende Lernmethode anwendbar, da bei dieser Lernmethode aufbereitete und speziell ausgewählte Trainings- sowie Testdaten vorliegen und die präsentierten Metriken solche Daten benötigen, um überhaupt ermittelt werden zu können (Machill, 2020).

Die erläuterten weiteren datentechnischen Maßnahmen wären ebenfalls generell nur für die überwachende Lernmethode anwendbar, da sie einerseits diejenigen Metriken verwenden, die nur mittels der überwachenden Lernmethode ermittelbar sind und andererseits weitere Methodiken aufweisen, welche nur mittels eines aufbereiteten und speziell ausgewählten Datensatzes durchführbar sind (Machill, 2020).

Dementsprechend benötigt man für KI-Systeme bzw. AES, die auf einer anderen Lernmethode, wie beispielsweise der unüberwachten Lernmethode, beruhen, andere Metriken sowie Techniken, um die Korrektheit sowie Fairness dieser AES gewährleisten zu können (Machill, 2020).

Hier empfiehlt es sich, anwendbare Techniken sowie Metriken zu entwickeln, da beispielsweise KI-Modelle, wie der aktuell im Trend liegende vorgestellte KI-Chatbot namens Chat-GPT, unter anderem mithilfe der unüberwachenden Lernmethode entwickelt worden ist und algorithmische Voreingenommenheiten aufweist.

Dass diese Voreingenommenheiten bei solch einem System schwieriger zu beheben sind, wird hiermit deutlich. Allerdings wird durch die Beliebtheit von Chat-GPT ebenfalls erkenntlich, dass dringend alternative Techniken sowie Metriken zu entwickeln sind.

## **10.5 Nachvollziehbarkeit trotz Intransparenz und Komplexität**

Bislang wurden zu fast allen Ursachen ausgleichende Maßnahmen vorgestellt, um die in AES möglicherweise aufkommenden algorithmischen Voreingenommenheiten zu mindern oder vorbeugend das Aufkommen dieser zu verhindern.

Nichtsdestotrotz wurden entgegenwirkende Maßnahmen bezüglich der großen Problematik der Intransparenz, welche aufgrund der hohen Komplexität der hier thematisierten AES zum Vorschein gelangt, nur teilweise erläutert.

Es wurde lediglich im Unterkapitel des Post-Processing illustriert, wie trotz eines komplexen und als Black-Box anzusehenden AES, nach dem initialen Training dieses System trotzdem algorithmische Voreingenommenheiten mithilfe von Nachverarbeitungsalgorithmen gemindert werden können. Dementsprechend existiert hiermit eine Möglichkeit, um algorithmische Verzerrungen trotz vorhandener Intransparenz und Komplexität zu mitigieren.

Allerdings fehlt die Erläuterung von Maßnahmen, um das Verhalten des KI-Systems, auch bei vorhandener Intransparenz und Komplexität, nachvollziehen zu können.

Diese *Nachvollziehbarkeit* wäre für die Umsetzung von voreingenommenheitsmindernden Methoden von großer Bedeutung, da somit besser identifiziert werden kann, auf Grundlage welcher konkreten Merkmale algorithmische Verzerrung im KI-System aufkommt.

Aufgrund der Tatsache, dass das Erlangen dieser Nachvollziehbarkeit durch Auflösen der Black-Box schwieriger ist als diese Black-Box so beizubehalten und mithilfe von Methodiken Rückschlüsse auf die Verwendung sowie Fairness der einzelnen Merkmale zu schließen, werden im Folgenden solche Methodiken vorgestellt (Johner, 2022).

Angefangen mit dem *Feature Effect*. Mithilfe dieser Methodik könnte man visualisiert zurückgeliefert bekommen, welche durchschnittlichen Auswirkungen ein ausgewähltes Merkmal auf die letztendlichen Vorhersagen des KI-Modells hat (Johner, 2022).

Weiter mit der *Permutation Feature Importance*. Diese Methodik könnte die Relevanz eines Merkmals bezüglich seiner Wichtigkeit für das zugrundeliegende Ergebnis ermitteln (Johner, 2022).

Hierfür würde man jedes Merkmal je einzeln modifizieren und überprüfen, ob die Ergebnisse gleichgeblieben sind oder sich verändert haben. Sollten sie gleichgeblieben sein, so hätte das modifizierte Merkmal keinen essenziellen Einfluss auf das zugrundeliegende Ergebnis, wohingegen eine Änderung des Ergebnisses bedeuten würde, dass das modifizierte Merkmal einen erheblichen Einfluss auf das Ergebnis hat (Johner, 2022).

Anhand der Schwere der Auswirkungen kann eine Rangliste erstellt werden, welche wiedergibt, welche Merkmale den größten und welche den geringsten Einfluss auf das Ergebnis des KI-Modells haben (Johner, 2022).

Als nächstes wird die Methodik *Shapley Values* vorgestellt. Diese kann dafür verwendet werden, um herauszufinden, wie das KI-Modell zu seiner jeweiligen Vorhersage gelangt ist. Hierfür wird das Zustandekommen der Prognose in die einzelnen einbezogenen Merkmale unterteilt, sodass verdeutlicht wird, welchen Anteil ein jeweiliges Merkmal zur Vorhersage beigetragen hat (Johner, 2022).

Mithilfe dieser Methodik lässt sich insbesondere gut erkenntlich machen ob geschützte Merkmale fälschlicherweise einen zu großen Anteil zur Ergebnisfindung beitragen (Johner, 2022).

Abschließend wird die Methodik *Counterfactual Explanations* näher erläutert. Mithilfe dieser kann erkenntlich gemacht werden, welche Merkmale geändert werden müssten, damit sich die Vorhersage des KI-Modells ebenfalls ändert. Hier sind mehrere mögliche Änderungskombinationen der Merkmale möglich, sodass die Vorhersage des KI-Modells sich auf Grundlage von verschiedenen Merkmalsänderungen umändern könnte (Johner, 2022).

Hiermit wurden Methodiken vorgestellt, mit welchen man das Verhalten des KI-Modells trotz vorhandener Intransparenz und Komplexität nachvollziehen kann.

Durch diese Methodiken erlangen die KI-Entwickler die Fähigkeit, bei einem Undurchschaubaren KI-System vorhandene algorithmische Voreingenommenheiten, die ansonsten eventuell unentdeckt geblieben wären, zu entdecken (Johner, 2022).

## **10.6 Unterstützende Tools**

Nimmt man das vorherige Kapitel hinzu, so konnten nun zu allen Ursachen und Problematiken ausgleichende Maßnahmen vorgestellt werden.

Mit diesen kann man aufkommende algorithmische Voreingenommenheiten mindern oder vorbeugend deren Aufkommen verhindern.

Nichtsdestotrotz sollten KI-Entwickler der Herausforderung des Identifizierens sowie Mindern dieser nicht allein gegenüberstehen, sondern sie sollten vielmehr vorhandene unterstützende Tools nutzen, um diese Herausforderung effizienter angehen zu können.

Dementsprechend werden im Folgenden drei, in der Literatur häufig zitierte, Unterstützungstools vorgestellt (Deoras, 2020; Gasser, 2019; Marques, 2021; Metwalli, 2021).

Anfangen mit dem *What-If-Tool* oder auch *was-wäre-wenn-Tool* von Google. Dieses Tool könnte zurückliefern, was für eine Vorhersage das KI-System treffen würde, wenn beispielsweise die Trainingsdatensätze verändert wären. Also liefert dieses Tool die Möglichkeit, in einer Untersuchung tiefgehende Anpassungen an den verwendeten Datensätzen vorzunehmen und dann zu prüfen, wie sich diese Änderungen auf das Verhalten sowie das Ergebnis des KI-Modells auswirken (Deoras, 2020; Metwalli, 2021).

Wichtig an dieser Stelle zu verstehen wäre, dass diese was-wäre-wenn-Änderungen nicht tatsächlich langfristig beispielsweise die Datensätze des Modells verändern würden, sondern diese Änderungen für lediglich eine Untersuchung gültig sind und danach wieder zurückgesetzt werden (Deoras, 2020; Metwalli, 2021).

Dementsprechend liefert dieses Tool mit seiner einfachen Benutzeroberfläche ein für die KI-Entwickler optimales Werkzeug, um zu untersuchen, wie ihr KI-System in verschiedensten Situationen agiert und welche möglichen algorithmischen Voreingenommenheiten dabei zum Vorschein gelangen (Deoras, 2020; Metwalli, 2021).

Ein weiteres Tool, welches zur Ermittlung von algorithmischer Verzerrung herangezogen werden kann, wäre *FairML*. Mithilfe des auf Python basierten Unterstützungstools kann geprüft werden, wie sich die Performance des KI-Systems abhängig von den übergebenen Eingabedaten verändert. Auf diese Weise können im Modell vorhandene algorithmische Voreingenommenheiten ermittelt werden (Metwalli, 2021).

Die bisher vorgestellten Tools würden das KI-Entwicklerteam zwar bei der Ermittlung von algorithmischer Verzerrung unterstützen, aber nicht wie diese ermittelten Voreingenommenheiten konkret gemindert werden können. Aus diesem Grund wird nun das Tool *AI Fairness 360* von IBM vorgestellt (Metwalli, 2021).

Dieses Tool liefert einerseits eine Vielzahl an Fairness-Metriken, um algorithmische Verzerrungen zu entdecken, und andererseits eine Reihe an Algorithmen, mit welchen diese Voreingenommenheiten mitigiert werden können. Hierbei werden bereits vorgestellte Techniken wie das Reweighting angewandt (Deoras, 2020; Metwalli, 2021).

In diesem Kapitel wurden somit unterstützende Werkzeuge vorgestellt, mit welchen das KI-Entwicklerteam effizienter in der Lage ist, algorithmische Voreingenommenheiten zu entdecken sowie zu mindern.

## **10.7 Überwachen & Überprüfen des KI-Systems**

Anhand des Kapitels „Ursachen algorithmischer Voreingenommenheit“ wurde deutlich, dass algorithmische Verzerrung und daraus resultierendes diskriminierendes Verhalten auf Grundlage vieler variierender Ursachen an verschiedenen Prozessschritten innerhalb des Entwicklungsprozesses eines KI-Systems bzw. AES entstehen kann.

Dementsprechend wäre zu den vorgestellten Maßnahmen zur Minderung algorithmischer Voreingenommenheit zusätzlich noch ein Entwicklungsprozess begleitendes kontinuierliches Überwachen sowie Überprüfen des Systems auf algorithmische Verzerrungen notwendig (De Montis, 2019).

Diese *Überwachung und Überprüfung* wird vom Entwicklerteam durchgeführt und beginnt in den ersten Entwicklungsschritten und greift bis hin zum produktiven Einsatz des Systems (De Montis, 2019).

Eines der ersten Überprüfungen hängt mit den Trainingsdaten zusammen. Es muss geprüft werden, ob anhand der ausgewählten Daten die Anforderungen an das System erreicht werden können (De Montis, 2019).

Außerdem muss untersucht werden, ob diese Daten Voreingenommenheiten enthalten. Dies wäre an dieser Stelle von besonderer Bedeutung, da Verzerrungen innerhalb der Trainingsdaten sich höchstwahrscheinlich in das Verhalten des Systems einschleusen würden (De Montis, 2019).

Zum einen wäre also der eigen erarbeitete Datensatz auf Voreingenommenheiten zu überprüfen, zum anderen und noch viel wichtiger wären zusätzlich hinzugezogene externe Datenquellen auf Verzerrungen zu prüfen, da man nicht davon ausgehen kann, dass diese ausgiebig geprüft wurden (De Montis, 2019).

Mit diesen überprüften Trainingsdaten wäre nun der Lernprozess des Modells anzugehen. Dieser muss ebenfalls überwacht und überprüft werden, da - selbst wenn die Trainingsdaten mit Sicherheit keine Voreingenommenheiten enthalten - das KI-Modell auf Grundlage der Art und Weise, wie dieses sein Verhalten erlernt, sich selbst algorithmische Verzerrungen aneignen könnte (De Montis, 2019).

Beispielsweise könnte es Zusammenhänge erkennen und in seine Entscheidungsfindung berücksichtigen, welche die Entwickler nicht berücksichtigt bzw. erkannt hatten, welche aber voreingenommenes Verhalten hervorrufen (De Montis, 2019).

Demensprechend muss das KI-System nach dem initialen Erlernen seines Verhaltens und ebenfalls nach den auf Grundlage der Feedback-Schleifen erfolgten Lerneinheiten auf algorithmische Verzerrung geprüft werden (De Montis, 2019).

Nichtsdestotrotz sollte das KI-System vor seiner produktiven Verwendung ausgiebig getestet werden. Das Testen sollte vielfältige Testfälle umfassen. Also sollten die Tests unterschiedliche Ethnien sowie Altersgruppen usw. abdecken mit dem Ziel zu identifizieren, ob bei diesen Ungleichbehandlungen vorliegen (De Montis, 2019).

Sollten welche vorliegen, so muss untersucht werden, welcher Ursache die jeweilige wahrscheinlich diskriminierende Ungleichbehandlung unterliegt (De Montis, 2019).

Ist das KI-System ausreichend getestet worden und befindet sich somit nun im produktiven Einsatz, so kann noch überprüft werden, ob die Endnutzer das System in dem Kontext verwenden, für welchen es entwickelt worden ist. Denn wie bereits erläutert, könnte eine Verwendung des Systems außerhalb seines Kontextes in einer Reduzierung der Genauigkeit der Prognosen und eventuell sogar in algorithmische Voreingenommenheit münden (De Montis, 2019).

Die durch diese Überwachungen und Überprüfungen entdeckten algorithmischen Verzerrungen sollten weiterhin beobachtet sowie weitestgehend gemindert werden. Hierfür kann das Entwicklerteam unter anderem diejenigen Maßnahmen durchführen, welche in diesem Kapitel vorgestellt wurden (De Montis, 2019; Gasser, 2019).

Weiterführend sollte das Entwicklerteam die Überwachung sowie Überprüfung des Systems nicht komplett allein durchführen. Stattdessen sollten weitere Personen mit unterschiedlicher

Ethnie, unterschiedlichem Geschlecht, unterschiedlichem Bildungsstand usw. hinzugezogen werden, um die Ergebnisse des Systems skeptisch zu betrachten und das Entwicklerteam somit auf mögliche unentdeckte algorithmische Verzerrungen aufmerksam zu machen (Gasser, 2019).

Dies stößt an die im Unterkapitel „Diversität“, erläuterten Vorteile von Vielfalt an.

Zusätzlich zur Überprüfung und Überwachung des Entwicklerteams können interdisziplinäre Teams im Auftrag von öffentlichen oder privaten Sektoren den Entwicklungsprozess untersuchen sowie kritisch hinterfragen (Gasser, 2019).

Mit diesem Kapitel konnte die Relevanz und Notwendigkeit des Überwachens sowie Überprüfens des Entwicklungsprozesses eines KI-Systems bzw. AES hervorgehoben werden. Außerdem wurde erläutert, an welchen Entwicklungsprozessschritten und eventuell ebenfalls mit welchen Maßnahmen das KI-Modell auf algorithmische Verzerrung geprüft werden könnte.

## **10.8 Rechtliche Maßnahmen**

Im Folgenden werden rechtliche Maßnahmen vorgestellt, welche implementiert werden könnten, um algorithmische Voreingenommenheit in KI-Systemen zu mindern.

Hier wird an die in Kapitel „Rechtliche Regelungen“ erörterten Gründe für die herrschende Präsenz verzerrter AES angeknüpft.

Zusätzlich wird auf die Wünsche der Bürger eingegangen. Diese wurden anhand einer Bevölkerungsumfrage evaluiert. In dieser Umfrage wurden sie darüber befragt, welche Maßnahmen sie sich bezüglich AES erhoffen (Kolleck & Orwat, 2020).

Die Umfrage ergab folgende gewünschten Maßnahmen:

- Sich zusätzlich zum Urteil des AES eine alternative Beurteilung einholen zu können (Kolleck & Orwat, 2020).



- Eine Auskunft darüber, welche personenbezogenen Daten das AES verwendet, wie diese verarbeitet werden und wie das AES zu seinem Urteil gelangt ist (Kolleck & Orwat, 2020).
- Eine Kennzeichnungspflicht, sodass gekennzeichnet wird, dass ein AES zur Entscheidungsfindung verwendet worden ist (Kolleck & Orwat, 2020).
- Dass diese AES von Dritten geprüft werden (Kolleck & Orwat, 2020).
- Einen Standard, welchen jedes AES erfüllen muss, um in der Praxis genutzt werden zu können. Ähnlich zum TÜV bei Fahrzeugen (Kolleck & Orwat, 2020).

Diese Maßnahmen spiegeln den Wunsch nach mehr eigenen Handlungsmöglichkeiten, mehr Transparenz sowie einheitlichen Sicherheitsstandards wider (Kolleck & Orwat, 2020).

### 10.8.1 **Transparenz**

Der Fokus liegt jedoch auf dem Wunsch nach mehr Transparenz. Die fehlende Transparenz AES und die möglichen Auswirkungen wurden in dieser Arbeit bereits ausgiebig erläutert (Kolleck & Orwat, 2020).

Ebenfalls ausführlich vorgestellt wurden mögliche Maßnahmen sowie Tools, um der Intransparenz von komplexen KI-Systemen bzw. AES entgegenwirken zu können.

Dementsprechend werden in diesem Unterkapitel zusätzlich noch rechtliche Maßnahmen vorgestellt, um der gesellschaftlichen Problematik der Intransparenz von AES entgegenwirken zu können.

Zuerst wäre wichtig hervorzuheben, dass eine komplette Offenlegung des Programmcodes sowie möglicherweise des zugrundeliegenden neuronalen Netzwerkes nicht das Ziel sein kann, um für mehr Transparenz zu sorgen. Einerseits wäre dies ein Verstoß gegen das Geschäftsgeheimnis des jeweiligen Unternehmens, andererseits würde dies den meisten Personen keine Klarheit verschaffen, da die technische Komplexität dieser Inhalte zu hoch ist für jemanden, der eventuell kein KI-Entwickler ist (Öktem, 2022).

Das Ziel hier wäre vielmehr, dass Betroffene über die Verwendung sowie den Zweck von AES informiert werden und Unternehmen zusätzlich Auskunft über die von ihnen getroffenen Maßnahmen zur Minderung von algorithmischer Diskriminierung liefern (Öktem, 2022; Kolleck & Orwat, 2020).

Also sollten Unternehmen durch eine Kennzeichnungspflicht dazu verpflichtet sein, den Betroffenen kenntlich zu machen, dass zur Festlegung der Entscheidung über sie ausschließlich oder unterstützend AES eingesetzt wurden (Öktem, 2022; Kolleck & Orwat, 2020).

Weiterführend sollten Betroffene ebenfalls darüber informiert werden, ob und wann Menschen zur Entscheidungsfindung des Urteils beitragen und was für Auswirkungen eine getroffene Entscheidung annehmen kann (Öktem, 2022).

Darüber hinaus muss den Betroffenen, insbesondere im Falle eines negativen Ergebnisses also z. B. im Falle einer Ablehnung für einen Kreditantrag, mitgeteilt werden, wie das System letztlich zu der finalen Entscheidung gekommen ist (Kolleck & Orwat, 2020).

Außerdem sollte das KI einsetzende Unternehmen einerseits verpflichtet sein, jegliche verwendeten Datensätze zu protokollieren, damit - falls Betroffene auf Verdacht von algorithmischer Diskriminierung hin eine Beschwerde einreichen - anhand dieser Protokolle einfacher mögliche Diskriminierungen ermittelt werden können (Kolleck & Orwat, 2020).

Andererseits sollte das KI entwickelnde Unternehmen das System so transparent sowie verständlich genug gestalten, dass Fachleute die Entscheidungsfindung des Modells nachvollziehen können. Dies wäre relevant, falls Dritte das AES beispielsweise auf seine menschenrechtlichen Auswirkungen prüfen müssten (Öktem, 2022).

Der Verpflichtung für mehr Transparenz kann ein Unternehmen mit Anwendung der in Kapitel 10.5 erläuterten Maßnahmen, für bessere Nachvollziehbarkeit und Transparenz von AES nachgehen.

### 10.8.2 Bekämpfung der Ohnmacht von Betroffenen

Betroffene von algorithmischer Diskriminierung wissen entweder gar nicht, dass sie diskriminiert worden sind oder sie wissen es, können jedoch aufgrund der Komplexität von AES, besonders aufgrund von Proxys, keine Indizien aufweisen, welche ihre Behauptung einer Diskriminierung bestätigen würde (Kolleck & Orwat, 2020).

Dementsprechend sollten Betroffene Unterstützung bei der Belegung einer möglicherweise vorliegenden algorithmischen Diskriminierung erhalten, um somit schnellen Rechtsschutz zu erlangen (Öktem, 2022; Kolleck & Orwat, 2020).

Eine Möglichkeit der Umsetzung dieser Unterstützung in Deutschland wäre mit einer Ausdehnung der Antidiskriminierungsstellen möglich. Diese Ausdehnung würde auf ihrem Schutzauftrag beruhen, wodurch sie stellvertretend für die möglicherweise diskriminierte Person „[...] stärker selbstständig investigativ vorgehen und rechtliche Schritte einleiten können“ (Kolleck & Orwat, 2020, S. 64) (Öktem, 2022; Kolleck & Orwat, 2020).

Alternativ zur Notwendigkeit des Betroffenen, die Indizien für Diskriminierung aufzuweisen, könnte das AGG die bereits vorhandene Beweislastumkehr für Diskriminierungsfälle erweitern, sodass wenn ein KI-System so komplex gestaltet ist, dass es als Black-Box-System anerkannt wird und eine Person von diesem System meint unfair behandelt worden zu sein, dies als Indiz für eventuell erfolgte Diskriminierung reiche und die Beweislastumkehr gelte (*Beweislastumkehr - Definition & Bedeutung im BGB bei Gewährleistung*, 2023; Kolleck & Orwat, 2020).

Diese Beweislastumkehr würde bedeuten, dass nun nicht mehr die betroffene Person beweisen muss, dass sie vom System diskriminiert worden ist, sondern das Unternehmen muss nun beweisen, dass das verwendete KI-System, die Person nicht diskriminiert hat (*Beweislastumkehr - Definition & Bedeutung im BGB bei Gewährleistung*, 2023; Kolleck & Orwat, 2020).

Mithilfe dieser Maßnahmen erhält eine betroffene Person die Möglichkeit zu handeln und Rechtsschutz zu erlangen, anstatt dem Urteil des AES Ohnmächtig ausgesetzt zu sein.

### 10.8.3 Kontrolle und Regulierung

Bevor sowie während ein Unternehmen ein AES zur Ermittlung von Entscheidungen verwendet, sollten jegliche menschenrechtliche Folgen überprüft werden. Hierbei sollte insbesondere untersucht werden, ob das KI-System algorithmische Diskriminierung aufweist oder aufweisen könnte (Öktem, 2022).

Diese menschenrechtlichen Auswirkungsüberprüfungen sollten vom Unternehmen selbst sowie von externen Dritten durchgeführt werden (Öktem, 2022).

Sollten diese Überprüfungen ein untragbares menschenrechtliches Risiko aufweisen, so wäre auf die Verwendung dieser AES zu verzichten (Öktem, 2022).

Damit diese Auswirkungsuntersuchungen sowie die Maßnahmen zur Minderung algorithmischer Voreingenommenheit einen ersthaften Charakter annehmen, sollten im Falle von Diskriminierung, einerseits Sanktionen an das Unternehmen verhängt werden und andererseits sollte immer eine „[...] juristische oder natürliche Person [...]“ (Öktem, 2022, S. 6) die Verantwortung für diese Vorfälle tragen (Öktem, 2022).

Außerdem, wie in Kapitel „Rechtliche Regelungen“ bereits angesprochen, sind die Auswirkungen von algorithmischer Voreingenommenheit unterschiedlich schwer. Dementsprechend könnte man anhand der Wahrscheinlichkeit des Diskriminierungseintritts zusammen mit der Schwere der Folgen des Eintritts, eine *Kritikalitätsstufe* für das System festlegen und anhand dieser Stufe könnten dann geeignete *Regulierungsmaßnahmen* durchgeführt werden (Kolleck & Orwat, 2020). Diese Maßnahmen umfassen:

- Ein komplettes oder teilweises Verbot der Nutzung des AES (Kolleck & Orwat, 2020)
- Die Notwendigkeit von „[...] Kontroll- und Transparenzpflichten [...]“ (Kolleck & Orwat, 2020, S. 65) (Kolleck & Orwat, 2020)
- Die Durchführung von Audits (Kolleck & Orwat, 2020)
- Keine Maßnahmen (Kolleck & Orwat, 2020)

Diese Kritikalitätsstufen sowie die jeweiligen Regulierungsmaßnahmen könnten sich an dem in Kapitel „Rechtliche Regelungen“ angeschnittenen Vorschlag für ein Gesetz über Künstliche

Intelligenz orientieren (*Europäisches Gesetz über Künstliche Intelligenz*, o. D.; Kolleck & Orwat, 2020).

Dieser Gesetzesentwurf schlägt *Risikoklassen* vor, in welche ein KI-System klassifiziert werden kann, um die Schwere der Auswirkungen dessen einstuft zu können (*Europäisches Gesetz über Künstliche Intelligenz*, o. D.).

Weiterführend hebt dieser Entwurf den Risikoklassen entsprechende gesetzliche Vorgaben und Regulierungen vor, um einer möglicherweise aufkommenden Diskriminierung proaktiv entgegenwirken zu können (*Europäisches Gesetz über Künstliche Intelligenz*, o. D.).

Eine weitere Regulierungsmaßnahme könnte in Form einer Selbstregulierung erfolgen. Hierbei könnten Unternehmen die KI-Systeme bzw. AES erstellen und/oder einsetzen, an eine Selbstverpflichtung gebunden werden, welche zu erfüllende Anforderungen umfasst wobei eine Nichterfüllung dieser in Offenlegungspflichten sowie eventuellen Sanktionen mündet (Kolleck & Orwat, 2020).

In diesem Kapitel wurden viele Kontroll-, Hilfs- und Regulierungsmaßnahmen erörtert, welche sich jedoch oftmals auf Möglichkeiten innerhalb Deutschlands bezogen haben.

Nichtsdestotrotz wurde bereits in den Ursachen die Notwendigkeit von expliziten sowie international geltenden Regulierungen bezüglich AES und algorithmischer Verzerrung hervorgehoben. Dementsprechend sollten die in dieser Arbeit thematisierten und weitere in den wissenschaftlichen Quellen zu lokalisierenden Möglichkeiten der Kontroll-, Hilfs- und Regulierungsmaßnahmen, europaweit oder sogar mithilfe der UN, interkontinental sowie einheitlich verankert werden (Kolleck & Orwat, 2020).

Mit diesen Maßnahmen sowie der Verankerung dieser im internationalen Raum könnte eine globale Minderung von algorithmischen Diskriminierungsfällen hervorgerufen werden.

An dieser Stelle wäre ebenfalls eine vom Autor dieser Arbeit ermittelte Handlungsempfehlung anzusprechen.

Das im AGG geschilderte Verbot der Verwendung von geschützten Merkmalen sollte aufgrund der Vorteile der Verwendung solcher Merkmale über Proxys (Kapitel 10.2.2) aufgehoben werden. Das AGG verbietet zwar ebenfalls die Verwendung von Proxys, doch wie in Kapitel 9.3.4

erwähnt, würde ein Proxy ein geschütztes Attribut ersetzen und ein entfernter Proxy würde von einem anderen Proxy ersetzt werden. Folglich müsste man nach aktueller gesetzlicher Regelung somit alle geschützten Attribute sowie Platzhalter entfernen. Dies würde jedoch in einem unbrauchbaren KI-System münden (Kapitel 9.3.4). Dementsprechend wäre die gesetzliche Vorgabe des AGG grundsätzlich technisch nicht umsetzbar, wodurch die vorteilhaftere Alternative das Erlauben der Verwendung von geschützten Merkmalen hervorgehoben wird. Wie in Kapitel 10.2.2 verdeutlicht, müssten durch die Einbeziehung von geschützten Merkmalen keine Proxys entfernt werden, da diese durch die direkte Verwendung der geschützten Attribute nicht mehr als Platzhalter fungieren. Aus diesen Gründen könnte das im AGG geschilderte Verbot im Ganzen aufgehoben werden.

# 11 Fazit & Ausblick

In diesem finalen Kapitel werden die wichtigsten Erkenntnisse dieser Arbeit zusammengefasst aufgeführt, wobei untersucht wird, ob das Ziel dieser Arbeit erreicht werden konnte. Darüber hinaus erfolgt ein Ausblick, in welchem darauf eingegangen wird, wie Forschende dieses Gebietes diese Arbeit weiterbehandeln könnten.

## 11.1 Fazit

Das Ziel dieser Arbeit war es, eine allumfassende Vorstellung sowie Lösung der Problematik von algorithmischer Voreingenommenheit in AES zu liefern.

Hierfür wurden zunächst einmal die für diese Arbeit relevanten Begrifflichkeiten vorgestellt und somit die Grundlage gelegt, mit der allumfassenden Darlegung dieses Themas zu beginnen.

Da sich diese mit den negativen Auswirkungen von KI-Systemen bzw. AES befasst, wurden vorher die wichtigsten Vorteile vorgestellt, die mit der Verwendung solcher Systeme einhergehen. Dies sollte illustrieren, dass dessen Einsatz mit den richtigen Intentionen erfolgt.

Anschließend folgt die umfassende Vorstellung des Themas mit der Fokussierung darauf, wie aufgrund des überwiegend unterbewussten Denkverhaltens unseres Gehirns, menschliche kognitive Verzerrungen unvermeidbar aufkommen und sich verfestigen. Dadurch konnte ausgiebig präsentiert werden, wie jeder Mensch kognitiven Voreingenommenheiten ausgesetzt ist, wobei diese einem oftmals nicht bewusst sind, da sie sich aus Effizienzgründen unterbewusst im Gehirn bilden.

Zusätzlich illustrieren die zu den kognitiven Verzerrungen präsentierten Beispiele, wie diese Verzerrungen in Voreingenommenheiten bezüglich der geschützten Merkmale eines Menschen münden könnten. Dies hebt hervor, dass diese Voreingenommenheiten nicht zwingend bewusst, sondern vielmehr durch die unterbewusst angeeigneten kognitiven Verzerrungen hervorkommen und sich somit jede Person solche Vorurteile aneignen könnte.

Ferner wurden diejenigen Voreingenommenheiten vorgestellt, die auf Grundlage der Verletzung der geschützten Merkmale hervorkommen können. Es wurde deutlich, dass diese Voreingenommenheiten insbesondere in einigen bestimmten Sektoren präsent sind, was die Grundlage für algorithmische Voreingenommenheit derjenigen AES darstellt, die in diesen Sektoren Anwendung finden oder fanden.

Im Anschluss wurden diejenigen voreingenommenen Sektoren vorgestellt, in welchen verzerrte KI-Systeme bzw. AES vermehrt Anwendung finden. Dies wurde mit dem Hintergrund durchgeführt, dass im weiteren Verlauf der Arbeit einige konkrete voreingenommene AES dieser Sektoren präsentiert werden.

Hierbei wurde illustriert, wie diese funktionieren, welche algorithmischen Voreingenommenheiten sie enthalten und welche Auswirkungen diese Voreingenommenheiten auf die Betroffenen dieser Systeme haben. Dies sollte hervorheben wie diskriminierend diese AES handeln können und wie präsent sowie komplex diese Thematik ist.

An dieser Stelle wurde das im Trend liegende KI-System Chat-GPT vorgestellt. Wobei verdeutlicht wurde, wie dessen algorithmischen Voreingenommenheiten eventuell noch verheerendere Auswirkungen annehmen könnten als die der vorangehenden Fallbeispiele.

Daraufhin wurden die Ursachen für algorithmische Voreingenommenheit präsentiert, wobei illustriert wurde, wie menschliche voreingenommene Handlungen in voreingenommenen Daten resultieren, welche den KI-Modellen zum Erlernen ihres Verhaltens zugeführt werden könnten, wodurch diese Modelle sich algorithmisch voreingenommenes Verhalten aneignen würden.

Aufgrund der Tatsache, dass menschliche voreingenommene Handlungen auf Grundlage von vorhandenen Voreingenommenheiten hervorkommen und diese Voreingenommenheiten wiederum auf Basis von kognitiven Verzerrungen entstehen, konnte verdeutlicht werden, dass



algorithmische Voreingenommenheit letzten Endes durch die unvermeidbaren kognitiven Verzerrungen eines Menschen hervorkommt.

Weiterführend wurde hervorgehoben, wie unausgewogene Datensätze in algorithmischen Verzerrungen münden können, wie sich diese Verzerrungen im KI-System verfestigen können und wie das Lernverhalten des Systems in weiteren algorithmischen Voreingenommenheiten münden kann.

Des Weiteren wurde verdeutlicht, wie Designentscheidungen und kognitive Verzerrungen der KI-Entwickler bzw. Datenwissenschaftler in algorithmischen Vorurteilen resultieren können.

Darüber hinaus wurde illustriert, wie selbst bei einem unvoreingenommenen System trotzdem diskriminierende Handlungen hervorkommen können, da die Endnutzer des Systems das System falsch verwenden oder die Ergebnisse dessen fehlinterpretieren.

Außerdem wurde hervorgehoben, wie komplex und intransparent diese Systeme werden können und welche Schwierigkeiten dies mit sich bringt. Abschließend wurde die nationale und internationale rechtliche Lage bezüglich algorithmischer Voreingenommenheit von AES niedergelegt, wobei hervorkam, dass unter anderem aufgrund der Intransparenz dieser Systeme und der fehlenden speziellen rechtlichen Regelungen, trotzdem noch voreingenommene AES existieren können.

Hiermit sollten die diversen Ursachen für das Aufkommen und Beibehalten von algorithmischer Voreingenommenheit hervorgehoben werden.

Im Anschluss wurden mögliche Lösungen zum Entgegenwirken von algorithmischer Voreingenommenheit vorgestellt.

Angefangen mit Wegen, wie menschliche Voreingenommenheiten erfasst und darüber hinaus bekämpft werden könnten. Hier wurde insbesondere die Relevanz eines diverseren Entwicklerteams, die Vorteile der Einbeziehung der Nutzer und Betroffene dieser KI-Systeme sowie die Notwendigkeit des Bewusstseins für diese Verzerrungen hervorgehoben.

Anschließend wurden datentechnische Maßnahmen erörtert, die der Bekämpfung von algorithmischer Voreingenommenheit dienen. Hier wurden primär zwei Ansätze vorgestellt: Zum

einen die Notwendigkeit eines ausgewogenen Datensatzes, zum anderen die Vorteile der Einbeziehung der geschützten Merkmale in die Entscheidungsfindung des AES.

Nachfolgend wurden statistische Metriken präsentiert, wobei an dieser Stelle mithilfe der Confusion-Matrix allgemeine statistische Metriken ermittelt werden konnten, mit welchen die Genauigkeit bestimmter KI-Modelle gemessen werden kann. Auf Grundlage dieser Metriken konnten Fairness-Metriken ermittelt werden, mit welchen die Fairness einiger KI-Modelle gemessen werden kann.

Auf Basis dessen konnten im weiteren Verlauf zusätzliche datentechnische Maßnahmen illustriert werden. Der Fokus lag hierbei auf Pre-Processing, In-Processing und Post-Processing bzw. Maßnahmen zur Minderung von algorithmischer Voreingenommenheit, welche während dieser Schritte durchgeführt werden könnten.

Daraufhin wurden Möglichkeiten vorgestellt, wie man trotz vorhandener Intransparenz und Komplexität die Entscheidungsfindung des AES nachvollziehen könnte. Dies würde den Entwicklern dabei helfen, das Verhalten dieser Systeme zu verstehen und Maßnahmen einzuleiten, um möglichem Fehlverhalten entgegenzuwirken.

Anschließend wurden verfügbare Tools präsentiert, welche die Entwickler bei der Bekämpfung von algorithmischer Voreingenommenheit assistieren könnten.

Des Weiteren wurde die Relevanz von Überwachungs- sowie Überprüfungsmaßnahmen hervorgehoben, da mithilfe dieser, algorithmische Voreingenommenheit besser entdeckt werden könnte. Es wurde empfohlen, diese während des gesamten Entwicklungsprozesses anzuwenden und somit von Anfang der Entwicklung bis hin zum produktiven Einsatz durchzuführen.

Zusätzlich wurden rechtliche Maßnahmen illustriert, die dabei helfen sollen, dass Betroffene von algorithmischer Diskriminierung rechtliche Gerechtigkeit erlangen.

Weiterhin kam die Relevanz von transparenteren AES durch eine Bevölkerungsumfrage zum Vorschein, sodass Maßnahmen zu Schaffung von transparenteren AES illustriert wurden.

Abschließend wurden Kontroll- und Regulierungsmaßnahmen dargelegt, mit welchen national sowie eventuell auch international für mehr rechtlichen Rückhalt gesorgt werden könnte. Die

konkrete Umsetzung dieser könnte eine globale Minderung von algorithmischen Diskriminierungsfällen hervorrufen.

Hiermit wurde auch der letzte Aspekt der umfassenden Erläuterung des Themas algorithmische Voreingenommenheit in AES aufgeführt.

Insbesondere wurde verdeutlicht, wie breit gefächert die Möglichkeiten der Aneignung sowie Minderung algorithmischer Voreingenommenheit sind und wie letzten Endes der Mensch als primäre Ursache sowie Lösung für diese Thematik fungiert.

Angesichts der allumfassenden Darlegung dieses Themas sowie der diversen Betrachtungsweise seiner Ursachen und Lösungen, kann das Ziel dieser Arbeit als erreicht angesehen werden.

## **11.2 Ausblick**

Aufgrund des immer weiter steigenden Trends der künstlichen Intelligenz und der voranschreitenden Entwicklung dieses Forschungsfeldes, liegt es nahe, dass AES und darüber hinaus die Problematik der algorithmischen Voreingenommenheit auch in der Zukunft eine präzente Thematik sein wird, sodass diese wissenschaftliche Ausarbeitung ein hochaktuelles Thema behandelt.

Diese Arbeit bietet eine umfassende Grundlage, um das Thema der algorithmischen Voreingenommenheit von AES nachzuvollziehen. Wie in der „Einleitung“ bereits erwähnt, wurde an einigen Stellen auf eine tiefgehende Erläuterung verzichtet, um den Rahmen dieser Arbeit nicht zu sprengen.

Folglich könnten weitere Forschungsarbeiten innerhalb dieses Themenspektrums diese Arbeit als Basis verwenden und bestimmte Teilaspekte dieser vertiefter erläutern.

Darüber hinaus bietet diese Arbeit die Grundlage, standardisierte Lösungsabläufe zu formulieren und den angeschnittenen rechtlichen Aspekt mit konkreten Richtlinien sowie Gesetzen zu spezifizieren.

Abschließend bietet die Arbeit ebenfalls die Möglichkeit, von weiteren Forschenden in seinen jeweiligen Kapiteln erweitert zu werden, sodass eine gemeinsam erarbeitete wissenschaftliche Ausarbeitung vorliegen könnte, welche als Grundlage für die Thematik von verschiedensten Interessenten verwendet werden kann. Diese Arbeit könnte auch in seiner jetzigen Verfassung als solch Grundlage verwendet werden, nichtsdestotrotz könnte das Erweitern seiner Kapitel in einer noch vollständigeren Wiedergabe der Thematik münden.

# Literaturverzeichnis

*Algorithmen und das Arbeitsmarktchancen-Modell des AMS.* (2019). Arbeitplus. Abgerufen 30. Juli 2023, von [https://arbeitplus.at/wordpress/wp-content/uploads/2020/06/2019-12\\_Position-Algorithmus-und-Segmentierung-1.pdf](https://arbeitplus.at/wordpress/wp-content/uploads/2020/06/2019-12_Position-Algorithmus-und-Segmentierung-1.pdf)

*Allgemeine Erklärung der Menschenrechte.* (2023). bmz. Abgerufen 30. Juli 2023, von <https://www.bmz.de/de/service/lexikon/allgemeine-erklaerung-der-menschenrechte-60138>

*Allgemeines Gleichbehandlungsgesetz (AGG) § 1 Ziel des Gesetzes.* (o. D.). Gesetze-im-internet. Abgerufen 30. Juli 2023, von [https://www.gesetze-im-internet.de/agg/\\_1.html](https://www.gesetze-im-internet.de/agg/_1.html)

Allhutter, D. (2021). Ein Algorithmus zur effizienten Förderung der Chancen auf dem Arbeitsmarkt?. *WISO - Wirtschafts- und Sozialpolitische Zeitschrift*, 2021(1), 82–97. [https://www.researchgate.net/publication/353347885\\_Ein\\_Algorithmus\\_zur\\_effizienten\\_Forderung\\_der\\_Chancen\\_auf\\_dem\\_Arbeitsmarkt\\_in\\_WISO\\_Wirtschafts-\\_und\\_Sozialpolitische\\_Zeitschrift\\_12021](https://www.researchgate.net/publication/353347885_Ein_Algorithmus_zur_effizienten_Forderung_der_Chancen_auf_dem_Arbeitsmarkt_in_WISO_Wirtschafts-_und_Sozialpolitische_Zeitschrift_12021)

*Amazon scrapped „sexist AI“ tool.* (2018). BBC. Abgerufen 30. Juli 2023, von <https://www.bbc.com/news/technology-45809919>

- Angwin, J. (2016). *Sample-COMPAS-Risk-Assessment-COMPAS-"CORE"*. ProPublica. Abgerufen 30. Juli 2023, von <https://www.documentcloud.org/documents/2702103-Sample-Risk-Assessment-COMPAS-CORE.html#document/p3/a296596>
- Anyoha, R. (2017, 28. August). The History of Artificial Intelligence. *Harvard*. Abgerufen 30. Juli 2023, von <https://sitn.hms.harvard.edu/flash/2017/history-artificial-intelligence/>
- Anzahl der Inhaftierten in den USA nach Ethnie im Jahr 2021*. (2023). Statista. Abgerufen 30. Juli 2023, von <https://de.statista.com/statistik/daten/studie/1124600/umfrage/inhaftierte-in-den-usa-nach-ethnie/>
- Aufstieg der Tech-Giganten*. (2021). Statista. Abgerufen 30. Juli 2023, von <https://de.statista.com/statistik/studie/id/83110/dokument/aufstieg-der-tech-giganten/>
- Balthasar, K. (2021, 19. Juni). Ankerheuristik im Online Marketing. *Waldhirsch*. Abgerufen 30. Juli 2023, von <https://waldhirsch.de/neuromarketing/ankerheuristik/>
- Bär, T. (2022). *Algorithmic Bias: Verzerrungen durch Algorithmen verstehen und verhindern*. Springer. <https://doi.org/10.1007/978-3-662-66315-8>
- Baum, J., & Villasenor, J. (2023, 8. Mai). *The politics of AI: ChatGPT and political bias*. Brookings. Abgerufen 30. Juli 2023, von <https://www.brookings.edu/articles/the-politics-of-ai-chatgpt-and-political-bias/>

Baumgartner, R., & Ernst, W. (2023). Künstliche Intelligenz in der Medizin? Intersektionale queerfeministische Kritik und Orientierung. *GENDER - Zeitschrift für Geschlecht, Kultur und Gesellschaft*, 15(1), 11–25. <https://doi.org/10.3224/gender.v15i1.02>

Beck, D. (2023, 16. Januar). *Gespräche führen mit ChatGPT: So lernt die KI von uns*. swr. Abgerufen 30. Juli 2023, von <https://www.swr.de/wissen/chatbots-wie-funktioniert-chat-gpt-100.html>

*Behinderung – gesetzliche Definitionen*. (o. D.). Studierendenwerke. Abgerufen 30. Juli 2023, von <https://www.studierendenwerke.de/themen/studieren-mit-behinderung/recht-politik-und-daten/behinderung-gesetzliche-definitionen>

*Behinderungen in der EU: Fakten und Zahlen*. (2022). Europa. Abgerufen 30. Juli 2023, von <https://www.consilium.europa.eu/de/infographics/disability-eu-facts-figures/>

Benson, B. (2016, 16. September). *You are almost definitely not living in reality because your brain doesn't want you to*. Quartz. Abgerufen 30. Juli 2023, von <https://qz.com/776168/a-comprehensive-guide-to-cognitive-biases>

Berger, V., & Schimansky, S. (2021, 14. August). *UNTER VERDACHT*. Forbes. Abgerufen 30. Juli 2023, von <https://www.forbes.at/artikel/unter-verdacht.html>

*Beweislastumkehr - Definition & Bedeutung im BGB bei Gewährleistung*. (2023, 12. Juli). Juraforum. Abgerufen 30. Juli 2023, von <https://www.juraforum.de/lexikon/beweislastumkehr>

*Bias*. (o. D.). Cambridge. Abgerufen 29. Juli 2023, von <https://dictionary.cambridge.org/dictionary/english/bias>

Black, D. A., Makar, H. R., Sanders, S. G., & Taylor, L. J. (2003). The earnings effects of sexual orientation. *Industrial & Labor Relations Review*, 56(3), 449–469.  
<https://doi.org/10.1177/001979390305600305>

Bocksch, R. (2022, 20. September). *Die USA haben die meisten Gefangenen weltweit*. Statista. Abgerufen 30. Juli 2023, von <https://de.statista.com/infografik/5560/gefangene-pro-100000-einwohner/>

Bönisch, J. (2010, 21. Mai). *Männer bevorzugt*. Süddeutsche. Abgerufen 30. Juli 2023, von <https://www.sueddeutsche.de/karriere/fuehrung-maenner-bevorzugt-1.889811>

Braswell, P. (2022, 19. Oktober). *This is the difference between racism and racial bias*. Fast-company. Abgerufen 30. Juli 2023, von <https://www.fastcompany.com/90796690/this-is-the-difference-between-racism-and-racial-bias>

Bremmer, M. (2022, 12. Dezember). *Auch ChatGPT hat ein Bias-Problem*. Computerwoche. Abgerufen 30. Juli 2023, von <https://www.computerwoche.de/a/auch-chatgpt-hat-ein-bias-problem,3613443>

*Brustkrebs (Mammakarzinom)*. (2022). Krebsdaten. Abgerufen 30. Juli 2023, von [https://www.krebsdaten.de/Krebs/DE/Content/Krebsarten/Brustkrebs/brustkrebs\\_node.html](https://www.krebsdaten.de/Krebs/DE/Content/Krebsarten/Brustkrebs/brustkrebs_node.html)



- Burch, K. (2022, 27. März). *Jeder Mensch hat unbewusste Vorurteile: Wie ihr sie überwinden könnt*. Business Insider. Abgerufen 30. Juli 2023, von <https://www.businessinsider.de/leben/jeder-mensch-hat-unbewusste-vorurteile-wie-ihr-sie-erkennen-und-ueberwinden-koennt/>
- Burns, E., Laskowski, N., & Tucci, L. (2023, 10. Juli). *artificial intelligence (AI)*. Techtarget. Abgerufen 30. Juli 2023, von <https://www.techtarget.com/searchenterpriseai/definition/AI-Artificial-Intelligence>
- Cashman, C. (2023, 27. Februar). *What is ChatGPT?*. Hp. Abgerufen 30. Juli 2023, von <https://www.hp.com/us-en/shop/tech-takes/what-is-chatgpt>
- ChatGPT*. (2023). Openai. Abgerufen 30. Juli 2023, von <https://chat.openai.com/>
- ChatGPT: Alles, was man wissen muss*. (o. D.). Moin ai. Abgerufen 30. Juli 2023, von <https://www.moin.ai/chatbot-lexikon/chatgpt-chatbot>
- Clickblogger. (2022, 29. September). *Vorteile von KI - zum Wohle der Gemeinschaft und Wirtschaft*. clickworker. Abgerufen 30. Juli 2023, von <https://www.clickworker.de/kunden-blog/vorteile-von-ki/>
- Common fairness metrics*. (o. D.). Fairlearn. Abgerufen 31. Juli 2023, von [https://fairlearn.org/main/user\\_guide/assessment/common\\_fairness\\_metrics.html](https://fairlearn.org/main/user_guide/assessment/common_fairness_metrics.html)
- Confirmation Bias: Warum der Bestätigungsfehler so gefährlich ist*. (o. D.). Barmer. Abgerufen 30. Juli 2023, von <https://www.barmer.de/gesundheitsverstehen/psyche/psychische-gesundheit/confirmation-bias-1128000>

Danks, D., & London, A. J. (2017). Algorithmic Bias in Autonomous Systems. *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*.

<https://doi.org/10.24963/ijcai.2017/654>

Dastin, J. (2018). Amazon scraps secret AI recruiting tool that showed bias against women.

*Reuters*. Abgerufen 30. Juli 2023, von <https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G>

De Montis, F. (2019). *Bias and its consequences in algorithms* [Fachhochschule St. Pölten].

<https://blog.felix.dm/content/files/2022/01/BAC1.pdf>

Deoras, S. (2020, 6. Oktober). *5 tools & frameworks that can clear bias from various da-*

*tasets*. Analytics India Magazine. Abgerufen 30. Juli 2023, von <https://analyticsindia-mag.com/5-tools-frameworks-that-can-clear-bias-from-various-datasets/>

Der Clustering Guide: Definition, Methoden und Beispiele. (o. D.). *Kobold AI*. Abgerufen 31.

Juli 2023, von <https://www.kobold.ai/clustering-guide/>

*Discrimination and Intolerance*. (o. D.). Council of Europe. Abgerufen 29. Juli 2023, von

<https://www.coe.int/en/web/compass/discrimination-and-intolerance>

Dörfler, C. (2023, 27. März). *Künstliche Intelligenz entdeckt Brustkrebs – vier Jahre, bevor*

*er ausbrach*. stylebook. Abgerufen 30. Juli 2023, von <https://www.stylebook.de/news/ki-brustkrebs-vorsorge>

Economic diversity and student outcomes at Harvard University. (2017). *The New York Times*. Abgerufen 30. Juli 2023, von <https://www.nytimes.com/interactive/projects/college-mobility/harvard-university>

*Einstellungsverfahren*. (o. D.). Polizei Hamburg. Abgerufen 29. Juli 2023, von <https://karriere-polizei.hamburg.de/einstellungsverfahren-3/>

Engelkamp, M. (o. D.). *Status Quo Bias (Status-quo-Verzerrung) – Definition & Beispiel*. DeltaValue. Abgerufen 30. Juli 2023, von <https://www.deltavalue.de/status-quo-bias/>

*Erwägungsgrund 71 - Profiling*. (o. D.). Datenschutz-Grundverordnung (DSGVO). Abgerufen 30. Juli 2023, von <https://dsgvo-gesetz.de/erwaegungsgruende/nr-71/>

*Europäisches Gesetz über Künstliche Intelligenz*. (o. D.). Deutsche Industrie- und Handelskammer. Abgerufen 30. Juli 2023, von <https://www.dihk.de/de/themen-und-positionen/wirtschaft-digital/dihk-durchblick-digital/europaeisches-gesetz-ueber-kuenstliche-intelligenz-63750>

Fam, P. (2023, 20. März). *ChatGPT: Was ist das, und was kann ich damit machen? Funktion, Nutzung und Kosten des Chatbots*. RedaktionsNetzwerk Deutschland. Abgerufen 30. Juli 2023, von <https://www.rnd.de/digital/chatgpt-was-ist-das-und-was-kann-ich-damit-machen-funktion-nutzung-und-kosten-des-chatbots-YYBITUVA6BG4NENF6VXWWTTJZQ.html>

*Feature Selection*. (2022). Alexander Thamm GmbH. Abgerufen 30. Juli 2023, von <https://www.alexanderthamm.com/de/data-science-glossar/feature-selection/>

- Fiegele, T. (2021). *Hired or rejected by an algorithm: the discourse of bias in algorithmic hiring* [Universität Innsbruck]. <https://digital.obvsg.at/ulbtirolhs/download/pdf/7061442?originalFilename=true>
- Fitzgibbons, L., Laskowski, N., & Botelho, B. (2022). *Data Scientist (Datenwissenschaftler)*. ComputerWeekly. Abgerufen 30. Juli 2023, von <https://www.computer-weekly.com/de/definition/Data-Scientist>
- Flores, A. W., Bechtel, K., & Lowenkamp, C. T. (2016). False Positives, False Negatives, and False Analyses: A Rejoinder to “Machine Bias: There’s Software Used Across the Country to Predict Future Criminals. And It’s Biased Against Blacks.”. *Federal Probation Journal*, 2016(2), 38–46. [https://www.uscourts.gov/sites/default/files/80\\_2\\_6\\_0.pdf](https://www.uscourts.gov/sites/default/files/80_2_6_0.pdf)
- Formen der Diskriminierung*. (2020, 23. April). Humanrights. Abgerufen 30. Juli 2023, von <https://www.humanrights.ch/de/ipf/menschenrechte/diskriminierung/diskriminierungsverbot-dossier/juristisches-konzept/formen-der-diskriminierung/>
- Forschner, A. (2021). *Malignes Melanom: schwarzer Hautkrebs*. Krebsgesellschaft. Abgerufen 30. Juli 2023, von <https://www.krebsgesellschaft.de/onko-internetportal/basis-informationen-krebs/krebsarten/hautkrebs/malignes-melanom-schwarzer-hautkrebs.html>
- Gasser, T. (2019). *Bias – A lurking danger that can convert algorithmic systems into discriminatory entities* [HAMK Hochschule für angewandte Wissenschaften Häme]. [https://www.theseus.fi/bitstream/handle/10024/167429/Gasser\\_Thea.pdf?sequence=2](https://www.theseus.fi/bitstream/handle/10024/167429/Gasser_Thea.pdf?sequence=2)

*Gedächtnistäuschung*. (o. D.). StudySmarter. Abgerufen 30. Juli 2023, von <https://www.studysmarter.de/schule/psychologie/grundlagendisziplinen-der-psychologie/gedaechtnistaeschungen/>

Glas, O. (2018, 14. September). Stereotype über Deutsche: Schnitzel, Lederhosen, Bier. *Deutsche Allgemeine Zeitung*. Abgerufen 30. Juli 2023, von <https://daz.asia/blog/stereotype-ueber-deutsche-schnitzel-lederhosen-bier/>

Greaton, T. (2019). *Was steckt hinter exponentiell steigendem Datenaufkommen?*. Nikko AM. Abgerufen 30. Juli 2023, von <https://www.nikkoam.de/articles/2020/whats-causing-the-exponential>

*Grundgesetz für die Bundesrepublik Deutschland Art 3*. (o. D.). Gesetze-im-internet. Abgerufen 30. Juli 2023, von [https://www.gesetze-im-internet.de/gg/art\\_3.html](https://www.gesetze-im-internet.de/gg/art_3.html)

Harpers, T. (2023, 22. März). *Mehrheit der Nutzer rechnet mit beruflichen Auswirkungen von ChatGPT*. RP ONLINE. Abgerufen 30. Juli 2023, von [https://rp-online.de/nrw/staedte/duisburg/chat-gpt-studie-bestaetigt-hohes-vertrauen-aber-hohe-fehlerquote\\_aid-87142131](https://rp-online.de/nrw/staedte/duisburg/chat-gpt-studie-bestaetigt-hohes-vertrauen-aber-hohe-fehlerquote_aid-87142131)

Hartmann, J., Schwenzow, J., & Witte, M. (2023). *The political ideology of conversational AI: Converging evidence on ChatGPT's pro-environmental, left-libertarian orientation*. <https://doi.org/10.48550/ARXIV.2301.01768>

- Hintze, A. (2016, 14. November). *Understanding the four types of artificial intelligence*. GovTech. Abgerufen 30. Juli 2023, von <https://www.govtech.com/computing/understanding-the-four-types-of-artificial-intelligence.html>
- Höger, S. (2023, 3. April). *Vereinte Nationen: Welche Länder sind Mitglieder?*. Augsburg Allgemeine. Abgerufen 30. Juli 2023, von <https://www.augsburger-allgemeine.de/politik/un-mitglieder-2023-liste-anzahl-der-laender-id62707546.html>
- Holzer, B. (2023, 9. März). *Maschinelle Mission*. FAZ; Frankfurter Allgemeine Zeitung GmbH. Abgerufen 30. Juli 2023, von <https://www.faz.net/aktuell/wissen/geist-soziales/chatgpt-chatbots-koennen-menschen-politisch-ueberzeugen-18720335.html>
- Human Rights*. (2023). United Nations. Abgerufen 30. Juli 2023, von [https://treaties.un.org/Pages/ViewDetails.aspx?src=TREATY&mtdsg\\_no=IV-3&chapter=4&clang=en](https://treaties.un.org/Pages/ViewDetails.aspx?src=TREATY&mtdsg_no=IV-3&chapter=4&clang=en)
- Hunts, A. (2019, 2. Januar). *Was sind künstliche neuronale Netze?*. rocketloop. Abgerufen 30. Juli 2023, von <https://rocketloop.de/de/blog/kuenstliche-neuronale-netze/>
- Is ChatGPT Woke And Biased? 16 examples that prove so*. (2023). Theinsaneapp. Abgerufen 30. Juli 2023, von <https://www.theinsaneapp.com/2023/02/chatgpt-woke-examples.html>
- Jahn, T. (2023, 19. Juni). ChatGPT: Was Sie über OpenAIs KI wissen sollten. *Handelsblatt*. Abgerufen 30. Juli 2023, von <https://www.handelsblatt.com/technik/it-internet/chatgpt-was-sie-ueber-openais-ki-wissen-sollten-/28941524.html>

- Jann, B., & Brunner, S. (2020). *Falsche Selbsteinschätzung hält Frauen von technischen Berufen fern*. Die Volkswirtschaft. Abgerufen 30. Juli 2023, von <https://dievolkswirtschaft.ch/de/2020/02/falsche-selbsteinschaetzung-haelt-frauen-von-technischen-berufen-fern/>
- Johner, C. (2022, 1. Februar). Interpretierbarkeit von KI: Blick in die Blackbox des maschinellen Lernens. *Johner Institut*. Abgerufen 30. Juli 2023, von <https://www.johner-institut.de/blog/regulatory-affairs/interpretierbarkeit-ki/>
- Johnson, A. (2023, 3. Februar). *Is ChatGPT partisan? Poems about trump and Biden raise questions about the AI bot's bias—here's what experts think*. Forbes. Abgerufen 30. Juli 2023, von <https://www.forbes.com/sites/ariannajohnson/2023/02/03/is-chatgpt-partisan-poems-about-trump-and-biden-raise-questions-about-the-ai-bots-bias-heres-what-experts-think/?sh=42fb1bed1371>
- Kamiran, F., & Calders, T. (2011). Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, 33(1), 1–33.  
<https://doi.org/10.1007/s10115-011-0463-8>
- Kilian. (2022, 8. Mai). Top 5 Vorteile beim Einsatz von künstlicher Intelligenz. *Kobold AI*. Abgerufen 30. Juli 2023, von <https://www.kobold.ai/ki-vorteile/>
- Kinderlähmung (Poliomyelitis)*. (2020). Sozialministerium. Abgerufen 30. Juli 2023, von [https://www.sozialministerium.at/Themen/Gesundheit/Uebertragbare-Krankheiten/Infektionskrankheiten-A-Z/Kinderl%C3%A4hmung-\(Poliomyelitis\).html](https://www.sozialministerium.at/Themen/Gesundheit/Uebertragbare-Krankheiten/Infektionskrankheiten-A-Z/Kinderl%C3%A4hmung-(Poliomyelitis).html)

Kolleck, A., & Orwat, C. (2020). *Mögliche Diskriminierung durch algorithmische Entscheidungssysteme und maschinelles Lernen – ein Überblick*. Büro für Technikfolgen-Abschätzung beim Deutschen Bundestag (TAB). <https://doi.org/10.5445/IR/1000127166>

*Korrelation als Kausalität*. (o. D.). Uni-leipzig. Abgerufen 30. Juli 2023, von <https://home.uni-leipzig.de/schreibportal/korrelation-als-kausaltaet/>

Koval, P. (2020, 29. Oktober). *Ansätze zur Überwindung von algorithmischer Voreingenommenheit*. Cassini. Abgerufen 30. Juli 2023, von <https://www.cassini.de/inspire/algorithmische-voreingenommenheit>

Künstliche Intelligenz Vor- und Nachteile. (2023). *Techparadies*. Abgerufen 30. Juli 2023, von <https://techparadies.de/kuenstliche-intelligenz-vorteile-nachteile/>

Langer, A. (2020, 21. August). *Datenschutzbehörde stoppt AMS-Algorithmus*. MeinBezirk. Abgerufen 30. Juli 2023, von [https://www.meinbezirk.at/c-wirtschaft/datenschutz-behoerde-stoppt-ams-algorithmus\\_a4200954?ref=curate](https://www.meinbezirk.at/c-wirtschaft/datenschutz-behoerde-stoppt-ams-algorithmus_a4200954?ref=curate)

Lawton, G. (2023, 7. Februar). *Datenvorverarbeitung (Data Preprocessing)*. Computer-Weekly. Abgerufen 30. Juli 2023, von <https://www.computerweekly.com/de/definition/Datenvorverarbeitung-Data-Preprocessing>

Linden, M. (2023, 3. Februar). *ChatGPT stellt Rekord für schnellwachsende Nutzerzahlen auf*. Golem. Abgerufen 30. Juli 2023, von <https://www.golem.de/news/kuenstliche-intelligenz-chatgpt-stellt-rekord-fuer-schnellwachsende-nutzerzahlen-auf-2302-171634.html>



- Litzel, N., & Luber, S. (2019, 25. April). *Was ist ein Expertensystem?*. BigData-Insider. Abgerufen 30. Juli 2023, von <https://www.bigdata-insider.de/was-ist-ein-expertensystem-a-819539/>
- Loeffner, A. (o. D.). Negativity Bias. *Psychologie des Glücks*. Abgerufen 30. Juli 2023, von <https://psychologie-des-gluecks.de/lexikon/negativity-bias/>
- Lohia, P. K., Ramamurthy, K. N., Bhide, M., Saha, D., Varshney, K. R., & Puri, R. (2018). *Bias mitigation post-processing for individual and group fairness*. <https://doi.org/10.48550/arXiv.1812.06135>
- Lohmeier, L. (2023). *Anzahl der Mitarbeiter von Amazon weltweit in den Jahren 2007 bis 2022*. Statista. Abgerufen 30. Juli 2023, von <https://de.statista.com/statistik/daten/studie/297593/umfrage/mitarbeiter-von-amazon-weltweit/>
- Machill, S. A. (2020). *Biased Artificial Intelligence Algorithmic Fairness and Human Perception of Biased AI* [NOVA IMS Schule für Informationsmanagement]. <https://run.unl.pt/bitstream/10362/109738/1/TGI0357.pdf>
- Mareike. (2019, 25. April). *Künstliche Intelligenz*. Tenmedia. Abgerufen 30. Juli 2023, von <https://www.tenmedia.de/de/glossar/kuenstliche-intelligenz>
- Marques, T. (2021). *Overcoming Algorithmic Bias: The role of Bias Awareness, Knowledge, and Minority Status on Human Decision-Making* [Katholische Universität Portugal]. [https://repositorio.ucp.pt/bitstream/10400.14/35384/1/152119110\\_Tiago%20Jos%C3%A1%20Freitas%20Marques\\_DPDFA.pdf](https://repositorio.ucp.pt/bitstream/10400.14/35384/1/152119110_Tiago%20Jos%C3%A1%20Freitas%20Marques_DPDFA.pdf)

- Maryfield, B. (2018). *Implicit Racial Bias*. Justice Research and Statistics Association. Abgerufen 30. Juli 2023, von <https://www.jrsa.org/pubs/factsheets/jrsa-factsheet-implicit-racial-bias.pdf>
- MasterClass. (2022, 6. März). *What Is Overconfidence Bias? 3 Types of Overconfidence Bias*. Masterclass. Abgerufen 30. Juli 2023, von <https://www.masterclass.com/articles/overconfidence-bias>
- Meschede, T. (2016). *Reiche Weiße, arme Schwarze*. Zeit. Abgerufen 30. Juli 2023, von [https://www.zeit.de/wirtschaft/2016-07/usa-weisse-schwarze-diskriminierung-oekonomische-unterschiede?utm\\_referrer=https%3A%2F%2Fwww.google.com%2F](https://www.zeit.de/wirtschaft/2016-07/usa-weisse-schwarze-diskriminierung-oekonomische-unterschiede?utm_referrer=https%3A%2F%2Fwww.google.com%2F)
- Mesevage, T. G. (2021, 24. Mai). *What is data preprocessing & what are the steps involved?*. *MonkeyLearn*. Abgerufen 30. Juli 2023, von <https://monkeylearn.com/blog/data-preprocessing/>
- Metwalli, S. A. (2021, 2. März). *5 Tools to Detect and Eliminate Bias in Your Machine Learning Models*. Towards Data Science. Abgerufen 30. Juli 2023, von <https://towardsdatascience.com/5-tools-to-detect-and-eliminate-bias-in-your-machine-learning-models-fb6c7b28b4f1>
- Mirza, S. A., & Rooney, C. (2018, 18. Januar). *Discrimination prevents LGBTQ people from accessing health care*. American Progress. Abgerufen 30. Juli 2023, von <https://www.americanprogress.org/article/discrimination-prevents-lgbtq-people-accessing-health-care/>

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society*, 3(2).

<https://doi.org/10.1177/2053951716679679>

Mohammad, S. (2021). *Bias in, bias out: A Study on Social Bias in Automated Decision-Making Algorithms* [Universität Utrecht]. [https://studentthe-](https://studenttheses.uu.nl/bitstream/handle/20.500.12932/1283/Mohammad%206280471%20.pdf?sequence=1&isAllowed=y)

[ses.uu.nl/bitstream/handle/20.500.12932/1283/Mohammad%206280471%20.pdf?sequence=1&isAllowed=y](https://studenttheses.uu.nl/bitstream/handle/20.500.12932/1283/Mohammad%206280471%20.pdf?sequence=1&isAllowed=y)

Muschter, R. (2023). *USA: Zugehörigkeit zu den Ethnien nach Selbstzuschreibung im Jahr*

2022. Statista. Abgerufen 30. Juli 2023, von <https://de.statista.com/statistik/daten/studie/166858/umfrage/ethnien-in-den-usa/>

Narkhede, S. (2018, 9. Mai). *Understanding confusion matrix*. Towards Data Science. Abge-

rufen 30. Juli 2023, von <https://towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62>

Naumann, M. (2023). *Was ist ChatGPT?*. ChatGPT auf deutsch nutzen und Alternativen. Abgerufen 30. Juli 2023, von <https://chatopenai.de/>

Öktem, R. (2022). *Algorithmische Entscheidungssysteme Menschenrechtliche Vorgaben und Entwicklungen auf internationaler Ebene*. Deutsches Institut für Menschenrechte.

[https://www.institut-fuer-menschenrechte.de/fileadmin/Redaktion/Publikationen/Information/Information\\_Algorithmische\\_Entscheidungssysteme.pdf](https://www.institut-fuer-menschenrechte.de/fileadmin/Redaktion/Publikationen/Information/Information_Algorithmische_Entscheidungssysteme.pdf)

- Oppermann, A. (2021, 26. August). *Accuracy, Precision, Recall, F1-Score und Specificity*. artemoppermann. Abgerufen 30. Juli 2023, von <https://artemoppermann.com/de/accuracy-precision-recall-f1-score-und-specificity/>
- Pawlik, V. (2022, 11. August). *Wie häufig hatten Sie schon das Gefühl benachteiligt zu sein, nur weil Sie eine Frau/ein Mann sind?*. Statista. Abgerufen 30. Juli 2023, von <https://de.statista.com/statistik/daten/studie/1052432/umfrage/umfrage-zu-diskriminierungserfahrungen-von-maennern-und-frauen-in-deutschland/>
- ProPublica — investigative journalism and news in the public interest*. (2023). ProPublica. Abgerufen 30. Juli 2023, von <https://www.propublica.org/>
- Pundy, D. (2019, 27. Januar). *Fall löst Kontroverse aus: Richter verurteilt Angeklagten mit Softwarehilfe zu Haftstrafe*. FOCUS. Abgerufen 30. Juli 2023, von [https://www.focus.de/panorama/welt/wenn-der-richter-mit-softwarehilfe-urteilt\\_id\\_10238880.html](https://www.focus.de/panorama/welt/wenn-der-richter-mit-softwarehilfe-urteilt_id_10238880.html)
- Raitner, M. (2018). *Eine kurze Geschichte der Digitalisierung*. raitner. Abgerufen 30. Juli 2023, von <https://raitner.de/2018/07/eine-kurze-geschichte-der-digitalisierung/>
- Raveling, J. (2022, 3. Mai). *Was ist ein neuronales Netz?*. Wfb-bremen. Abgerufen 30. Juli 2023, von <https://www.wfb-bremen.de/de/page/stories/digitalisierung-industrie40/was-ist-ein-neuronales-netz>

- Räz, T. (2022). COMPAS: zu einer wegweisenden Debatte über algorithmische Risikobeurteilung. *Forensische Psychiatrie Psychologie Kriminologie*, 16(4), 300–306.  
<https://doi.org/10.1007/s11757-022-00741-9>
- Rosso, C. (2018, 6. Februar). The human bias in the AI machine. *Psychology Today*. Abgerufen 30. Juli 2023, von <https://www.psychologytoday.com/intl/blog/the-future-brain/201802/the-human-bias-in-the-ai-machine>
- Rudin, C., Wang, C., & Coker, B. (2020). The age of secrecy and unfairness in recidivism prediction. *Harvard Data Science Review*, 2(1).  
<https://doi.org/10.1162/99608f92.6ed64b30>
- Rütten, F. (2018, 30. Januar). *Compas bei Gericht: Wo ein Algorithmus bestimmt, ob jemand rückfällig wird*. STERN. Abgerufen 30. Juli 2023, von <https://www.stern.de/panorama/verbrechen/compas-bei-gericht--wo-ein-algorithmus-bestimmt--ob-jemand-rueckfaellig-wird-7843206.html>
- Schick, U. (2018, 20. März). *Was ist künstliche Intelligenz?*. Sap. Abgerufen 30. Juli 2023, von <https://news.sap.com/germany/2018/03/was-ist-kuenstliche-intelligenz/>
- Schulz, L. M. (2020, 7. März). *Bloß keine Vorurteile?*. Spektrum. Abgerufen 30. Juli 2023, von <https://www.spektrum.de/news/bloss-keine-vorurteile/1708616>
- Schutz von Menschenrechten*. (o. D.). Tatsachen über Deutschland. Abgerufen 30. Juli 2023, von <https://www.tatsachen-ueber-deutschland.de/de/frieden-und-sicherheit/schutz-von-menschenrechten>

*Schwarze häufiger unschuldig verurteilt als Weiße.* (2022, 28. September). Migazin. Abgerufen 30. Juli 2023, von <https://www.migazin.de/2022/09/28/us-studie-schwarze-haeufiger-unschuldig-verurteilt-als-weisse/>

Six-Materna, I., & Six, B. (o. D.). *Stereotype*. Spektrum. Abgerufen 30. Juli 2023, von <https://www.spektrum.de/lexikon/psychologie/stereotype/14836>

Stadler, M.-L. (2020, 24. Februar). In 5 Schritten zu Ihrem KI-Modell. *mindsquare*. Abgerufen 30. Juli 2023, von <https://mindsquare.de/fachartikel/in-5-schritten-zu-ihrem-ki-modell/>

Stappert, L. (2022, 23. Juni). Was sind kognitive Verzerrungen? Definition und Beispiele. *Hubspot*. Abgerufen 30. Juli 2023, von <https://blog.hubspot.de/marketing/kognitive-verzerrung>

Swaminathan, N. (2022, 25. Januar). *Understanding Algorithmic Bias*. Medium. Abgerufen 30. Juli 2023, von <https://medium.com/the-research-nest/understanding-algorithmic-bias-18b9d1b935ca>

Thakur, N. K. (2021, 9. Juli). *Less Known Applications of k-Means Clustering — Dimensionality Reduction, Anomaly Detection and Data Representation*. Medium. Abgerufen 30. Juli 2023, von <https://medium.com/analytics-vidhya/less-known-applications-of-k-means-clustering-dimensionality-reduction-anomaly-detection-and-908f4bee155f>

*Verarbeitung besonderer Kategorien personenbezogener Daten.* (o. D.). Datenschutz-Grundverordnung (DSGVO). Abgerufen 30. Juli 2023, von <https://dsgvo-gesetz.de/art-9-dsgvo/>

Verma, S., & Rubin, J. (2018). Fairness definitions explained. *Proceedings of the International Workshop on Software Fairness*. <https://doi.org/10.1145/3194770.3194776>

*Verteilung der Körpergrößen nach Geschlecht im Jahr 2006.* (2006). Statista. Abgerufen 30. Juli 2023, von <https://de.statista.com/statistik/daten/studie/1825/umfrage/koerpergroesse-nach-geschlecht/>

Wagner, B., Lopez, P., Cech, F., Grill, G., & Sekwenz, M.-T. (2020). Der AMS-Algorithmus: Transparenz, Verantwortung und Diskriminierung im Kontext von digitalem staatlichem Handeln. *Zeitschrift für kritik - recht - gesellschaft*, 2, 191. <https://doi.org/10.33196/juridikum202002019101>

*Was kostet ein Studium im Ausland?.* (o. D.). Bachelor-and-more. Abgerufen 30. Juli 2023, von <https://www.bachelor-and-more.de/magazin/ausland/studiengebuehren-ausland/>

*What is Noise in Data Mining.* (o. D.). Javatpoint. Abgerufen 30. Juli 2023, von <https://www.javatpoint.com/what-is-noise-in-data-mining>

*Wie groß ist das Internet?.* (2015). FOCUS. Abgerufen 30. Juli 2023, von [https://www.focus.de/digital/multimedia/glasers\\_modernste\\_zeiten/wie-gross-ist-das-internet-alle-daten-des-planeten\\_id\\_1826941.html](https://www.focus.de/digital/multimedia/glasers_modernste_zeiten/wie-gross-ist-das-internet-alle-daten-des-planeten_id_1826941.html)

Wie künstliche Intelligenz die Softwareentwicklung revolutioniert. (o. D.). *Alpisoft*. Abgerufen 30. Juli 2023, von <https://alpisoft.net/blog/programmieren-mit-chatgpt-wie-kuenstliche-intelligenz-die-softwareentwicklung-revolutioniert>

Wissgott, P. (2019). The failure of Amazon's AI recruiting engine. *Linkedin*. Abgerufen 30. Juli 2023, von <https://www.linkedin.com/pulse/failure-amazons-ai-recruiting-engine-philipp-wissgott/>

Wolf, L. (2023, 1. Februar). *The damage done to the credibility of AI by ChatGPT engineers building in political bias is irreparable*. *Twitter*. Abgerufen 30. Juli 2023, von <https://twitter.com/LeighWolf/status/1620744921241251842>

Wuttke, L. (2022a, 21. Juni). *Was ist Supervised Learning (Überwachtes Lernen)?*. *datasolut*. Abgerufen 30. Juli 2023, von <https://datasolut.com/wiki/supervised-learning/>

Wuttke, L. (2022b, 21. Juni). *Was ist Unsupervised Learning (Unüberwachtes Lernen)?*. *datasolut*. Abgerufen 30. Juli 2023, von <https://datasolut.com/wiki/unsupervised-learning/>

Wuttke, L. (2023a, 24. Mai). *Machine Learning: Definition, Algorithmen, Methoden und Beispiele*. *datasolut*. Abgerufen 30. Juli 2023, von <https://datasolut.com/was-ist-machine-learning/>

Wuttke, L. (2023b, 24. Mai). *Künstliche Neuronale Netzwerke: Definition, Einführung, Arten und Funktion*. *datasolut*. Abgerufen 30. Juli 2023, von <https://datasolut.com/neuronale-netzwerke-einfuehrung/>




Zhang, B. H., Lemoine, B., & Mitchell, M. (2018). Mitigating unwanted biases with adversarial learning. *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. <https://doi.org/10.1145/3278721.3278779>

## Erklärung zur selbstständigen Bearbeitung einer Abschlussarbeit

Hiermit versichere ich, dass ich die vorliegende Arbeit ohne fremde Hilfe selbständig verfasst und nur die angegebenen Hilfsmittel benutzt habe. Wörtlich oder dem Sinn nach aus anderen Werken entnommene Stellen sind unter Angabe der Quellen kenntlich gemacht.

\_\_\_\_\_

Ort                      Datum                       Unterschrift im Original