Bundesamt
für Sicherheit in der
Informationstechnik

Deutschland
Digital•Sicher•BSI•

Bias in der künstlichen Intelligenz

Änderungshistorie

Version	*Datum*	*Name*	*Beschreibung*
1.0	01.07.2025	Dr. Jonas Ditz, Elmar Lichtmeß	Initiale Veröffentlichung
1.1	01.08.2025	Dr. Jonas Ditz, Elmar Lichtmeß	Englischer Begriff des Schutzziels Verfügbarkeit in der Einleitung korrigiert

Tabelle 1: Versionsverlauf

Bundesamt für Sicherheit in der Informationstechnik
Postfach 20 03 63
53133 Bonn
E-Mail: ki-kontakt@bsi.bund.de
Internet: https://www.bsi.bund.de

Zusammenfassung

Bias ist eine tiefgreifende Problematik für Systeme der Künstlichen Intelligenz, die eine Gefährdung für den sicheren Einsatz von solchen Systemen darstellen kann. Der richtige Umgang mit Bias ist komplex und erfordert eine umfassende Beschäftigung mit der Thematik. Diese Publikation soll Entwickelnden, Anbietenden und Betreibenden von KI-Systemen eine erste Einführung in die Bias-Thematik geben.

Bias-Arten weisen eine hohe Diversität auf und können in unterschiedlichen Phasen des Lebenszyklus eines KI-Systems auftreten. Eine Auswahl der relevantesten Bias-Arten sowie eine Zuordnung zur Phase des Lebenszyklus, in der diese auftreten, wird dargestellt. Die Detektion von Bias ist sowohl in Daten als auch in bereits trainierten KI-Modellen möglich. Eine Auswahl an qualitativen und quantitativen Detektionsmöglichkeiten wird präsentiert. Anschließend gibt das Manuskript eine Übersicht über Methoden, die verwendet werden können, um Bias in einem KI-System zu verhindern. Diese sind aufgeteilt in Prä-, In- und Postprozessierungsmethoden, abhängig von dem jeweiligen Zeitpunkt, zu dem sie angewendet werden können. Zum Abschluss wird die Interaktion zwischen Bias und Cybersicherheit diskutiert.

Das BSI fordert von Entwickelnden, Anbietenden und Betreibenden von KI-Systemen folgende Handlungen:

- **Wissen hilft bei der Lösung des Problems.** Es muss ein ausreichender Informationsaufbau zur Bias-Thematik stattfinden, um entscheiden zu können, welche Arten für die eigene Anwendung relevant sind. (Kapitel 2)
- **Zuständigkeiten müssen geklärt sein.** Für verwendete Datensätze und KI-Systeme muss eine für Bias zuständige Person benannt werden. Diese Person muss in der Lage sein, potentielle Bias zu identifizieren und geeignete Gegenmaßnahmen vorzuschlagen.
- **Bekämpfung von Bias beginnt bei den Daten.** Nach Möglichkeit müssen organisatorische und technische Maßnahmen bei der Datenerhebung etabliert werden, die potentiellen Bias in den Daten reduzieren. (Kapitel 3)
- **Ziel muss die Minimierung von unerwünschten Bias in KI-Modellen sein.** Prä- und Inprozessierungsmaßnahmen müssen – falls anwendbar – vorrangig verwendet werden. Vor der Anwendung von vortrainierten KI-Modellen (beispielsweise KI-Systeme mit allgemeinem Verwendungszweck) muss die Notwendigkeit für Postprozessierungsmaßnahmen geprüft werden (Kapitel 4)
- **Bei Bias heißt es am Ball bleiben.** Bias-Detektion und -Mitigation muss als Prozess verstanden werden und fester Bestandteil jeder Phase des Produktzyklus eines KI-Modells sein.

Inhalt

1 Einleitung 5
2 Bias-Arten 7
2.1 Bias bei der Datenerhebung 7
2.2 Bias bei der Entwicklung von KI-Systemen 9
2.3 Bias bei der Nutzung eines KI-Systems 10
2.4 Weitere Begrifflichkeiten 12
3 Bias-Detektion 13
3.1 Detektion von Bias in Daten 13
3.2 Detektion von Bias in KI-Modellen 15
4 Bias-Mitigation 19
4.1 Präprozessierungsmethoden 19
4.2 Inprozessierungsmethoden 22
4.3 Postprozessierungsmethoden 24
5 Bias und Cybersicherheit 27
5.1 Auswirkung auf das Schutzziel Vertraulichkeit 27
5.2 Auswirkung auf das Schutzziel Integrität 27
5.3 Auswirkung auf das Schutzziel Verfügbarkeit 28
6 Fazit 29
Literaturverzeichnis 30

1 Einleitung

Durch den sich immer weiter ausbreitenden Einsatz von Methoden der Künstlichen Intelligenz (KI) rücken verstärkt Themen abseits von Fragen zu technischen Möglichkeiten und Limitationen der Technologie in den Fokus. Ein Thema wird dabei häufig und kontrovers diskutiert: Die Ungleichbehandlung von Subpopulationen innerhalb des Datenraumes[1] durch KI-Systeme (Angwin, et al., 2016) (Buolamwini, et al., 2018) (Dastin, 2022) (ISO/IEC TR 24027, 2021). In diesem Zusammenhang wird gerne der Begriff Bias[2] eingebracht. Dieser Begriff beschreibt die genannte, durch Verzerrungen in den Daten ausgelöste Ungleichbehandlung durch KI-Systeme[3]. Solche Verzerrungen können beispielsweise eine übermäßige Betonung von problematischen Mustern innerhalb der Datenmerkmale sein. Auch eine fehlende oder zu geringe Repräsentation von bestimmten Subpopulationen wäre eine solche Verzerrung.

Selbst KI-Systeme, die in bester Absicht sowie nach dem neuesten Stand der Technik erstellt werden, können von Bias betroffen sein. Denn Bias ist häufig bereits in den Daten vorhanden. Datensammlungen beinhalten immer menschlichen Einfluss und sind gefärbt durch zahlreiche technische, wirtschaftliche, rechtliche und soziale Entscheidungen. Aber auch Designentscheidungen bei der Konzeption eines KI-Modells sowie beim Trainings- und Auswahlprozess können zu Bias im Verhalten der resultierenden KI-Modelle führen. Um eine informierte Einschätzung zur Gefahr von durch Bias beeinflussten Ausgaben eines KI-Systems treffen zu können, reicht also kein punktuelles Vorgehen. Diese Gefahrenabschätzung muss vielmehr Teil des gesamten Lebenszyklus eines KI-Systems sein, beginnend bei der Datenerhebung über die Modellkonzeption und dem Modelltraining bis zum Einsatz eines KI-System mit evtl. anfallender Interaktion mit Nutzenden.

Die möglichen Auswirkungen von Bias können weitreichend sein. Für Nutzende von KI-Systemen kann eine durch Bias ausgelöste Ungleichbehandlung zu diskriminierenden Ergebnissen führen, wodurch Nutzenden beispielsweise unberechtigterweise Zugang zu Ressourcen oder Gelegenheiten verwehrt wird. Aber auch Unternehmen, die KI-Systeme in ihren Geschäftsprozessen oder Produkten einsetzen, können durch Bias-gefärbtes Verhalten geschädigt werden. Beispielsweise wenn eine Ungleichbehandlung durch das eigene Produkt zu schadensersatzpflichtigen Situationen führt. Oder wenn Bias zu einem unerwarteten Verhalten führt, welches die Geschäftsprozesse stört.

Des Weiteren bedeutet Bias auch eine Gefahr für die IT-Sicherheit im Allgemeinen. Wie bereits erwähnt, kann Bias zu unerwarteten Verhalten von KI-Systemen mit Auswirkungen für Subpopulationen innerhalb des Datenraumes führen. Hierdurch können neue Angriffsvektoren auf IT-Systeme entstehen, wenn maliziöse Akteure gezielt diese unerwarteten Verhaltensweisen ausnutzen. Wenn KI-Systeme in sicherheitsrelevanter Infrastruktur eingesetzt werden, wird Bias außerdem zu einem generellen Sicherheitsproblem. Beispielsweise könnte der Einsatz einer biometrischen Zugangskontrolle auf Basis von KI-

[1] Der Begriff Datenraum umschreibt vereinfacht gesagt das Einsatzgebiet eines KI-Systems. Normalerweise wird die Ungleichbehandlung durch KI-Systeme im Kontext von menschlichen Nutzenden diskutiert. In diesem Fall würde der Datenraum aus Menschen bestehen und Subpopulationen wären in diesem Fall beispielsweise verschiedene Geschlechtsidentitäten oder Ethnizitäten. Die Problematik besteht jedoch auch für beliebige Datenräume und kann dort ebenfalls zu unerwünschten oder bedenklichen Verhalten führen.

[2] In der deutschen Literatur werden teilweise die Wörter „Schiefe“ und „Verzerrung“ äquivalent zum Wort „Bias“ verwendet.

[3] In diesem Manuskript wird explizit der sogenannte *Inductive Bias* (dt.: Induktive Verzerrung) ausgenommen. Hierbei handelt es sich um grundlegende Annahmen, die gemacht werden müssen, um aus Trainingsdaten lernen zu können. *Inductive Bias* ist eine Grundvoraussetzung für maschinelles Lernen und künstliche Intelligenz.

Modellen zu einem Sicherheitsvorfall führen, wenn das KI-Modell Menschen einer bestimmten Ethnizität Bias-bedingt nicht adäquat von Menschen einer anderen Ethnizität unterscheiden kann. Im schlimmsten Fall könnten hierdurch unberechtigte Personen Zugang zu als sicherheitsrelevant eingestuften Orten oder Daten erhalten. In anderen Worten, Bias kann sich negativ auf drei der wichtigsten IT-Schutzziele auswirken, die im „CIA-Dreieck" zusammengefasst sind: Vertraulichkeit (Confidentiality), Integrität (Integrity) und Verfügbarkeit (Availability). Aus diesen Gründen ist es von gehobener Bedeutung, Bias bei Entwicklung und Einsatz von KI mitzudenken.

Diese Publikation richtet sich an Anbieter[4], Betreiber[4], Bevollmächtigte[4] und Importeure[4] von KI-Systemen. Sie soll eine Einführung in den Themenkomplex Bias geben und wichtige Bias-Arten vorstellen. Des Weiteren werden Ansätze zur Detektion von Bias vorgestellt, die genutzt werden können, um das Auftreten von Bias in Daten und KI-Modellen abschätzen zu können. Abschließend werden verschiedene Methoden vorgestellt, die zum Ziel haben, Bias im Verhalten von KI-Modellen zu minimieren. Die Publikation betrachtet hierbei Bias aus einer statistischen, technischen Sicht. Es wird keine Diskussion zur übergeordneten Fairness-Thematik durchgeführt. Generell kann festgehalten werden, dass eine Ungleichbehandlung durch KI-Systeme leicht zu Voreingenommenheit und Diskriminierung führen kann, wenn solche Systeme auf Menschen angewendet werden. Die interessierten Lesenden seien hier auf die einschlägige Literatur verwiesen (Deutscher Ethikrat, 2023) (Barocas, et al., 2023) (Strotherm, et al., 2024).

[4] Entsprechend der Definition dieser Begriffe in der Verordnung über künstliche Intelligenz (KI-VO).

2 Bias-Arten

Die Bias Arten sind vielfältig, und die Ursachen und Risiken von Bias in KI-Systemen sind vielschichtig und tief in den technischen Aspekten der KI-Entwicklung verankert. In der Fachliteratur werden zahlreiche verschiedene Arten von Bias definiert. Diese unterscheiden sich in ihrer Spezifität und überschneiden sich teilweise.

Im Folgenden wird eine Auswahl an Bias-Arten vorgestellt. Diese ist sortiert nach den Lebenszyklusphasen, in denen die Bias-Arten entstehen bzw. vorkommen können. Der Lebenszyklus eines KI-Systems kann in Abbildung 1 eingesehen werden. Zum Abschluss des Kapitels werden weitere Begrifflichkeiten vorgestellt, die für interessierte Lesende vorteilhaft sein können, um eine bessere Einführung in die Bias-Thematik zu gewährleisten.

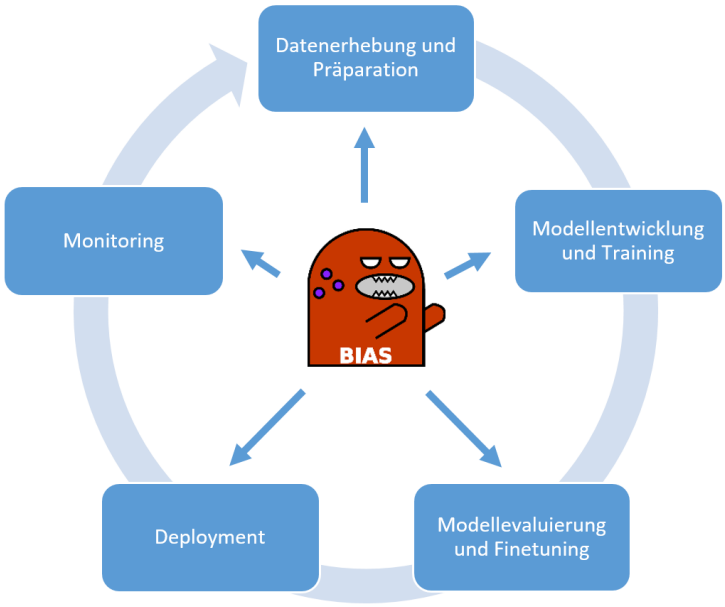

Abbildung 1: Bias steckt in jeder Phase des Lebenszyklus eines KI-Systems. Dieses Kapitel stellt Bias-Arten aus verschiedenen Phasen vor. Im Unterkapitel 2.1 werden Bias-Arten aus der Phase „Datenerhebung und Präparation“ vorgestellt. Unterkapitel 2.2 zeigt Bias-Arten aus den Phasen „Modellentwicklung und Training“ und „Modellevaluierung und Finetuning“ auf. Abschließend werden im Unterkapitel 2.3 Bias-Arten aus den Phasen „Deployment“ und „Monitoring“ vorgestellt.

2.1 Bias bei der Datenerhebung

Zuerst werden Bias-Arten betrachtet, die bereits bei der Datenerhebung und -präparation auftreten können.

A1 Historischer Bias (Historical Bias)

Historischer Bias tritt durch Datenverzerrungen auf, die zustande kommen, wenn Daten Sichtweisen und Konzepte widerspiegeln, die in einer modernen Welt nicht mehr zeitgemäß sind. Indem beim Konzipieren und Trainieren eines KI-Systems solche Daten verwendet werden, können KI-Systeme statistische Zusammenhänge auf Basis dieser „veralteten“ Sichtweisen lernen. Insofern können bereits bestehende Vorurteile und ungerechtfertigte Assoziationen in die Entscheidungsfindung des KI-Systems gelangen – unabhängig davon, wie perfekt Datenerhebung und Stichproben durchgeführt werden.

Beispiel: Eine KI, die zur Unterstützung bei der Bewerberauswahl eines Unternehmens eingesetzt werden soll, wird mit älteren Daten trainiert. Diese Daten repräsentieren aber noch

eine Zeit, in der im Unternehmen deutlich mehr Männer als Frauen eingestellt wurden. Dadurch wird diese Präferenz ungewollt auch in zukünftigen Auswahlentscheidungen fortgesetzt und eine übermäßige Repräsentation von Männern zementiert.

A2 Repräsentationsbias (Representation Bias)

Von einem Repräsentationsbias wird gesprochen, wenn Teile bzw. Feinstrukturen von Subpopulationen in den Trainingsdaten fehlen oder unterrepräsentiert werden und hierdurch Ungleichbehandlung entsteht.

Der Bias kann durch eine Schieflage bei der zugrunde liegenden Verteilung der Daten entstehen. Aber auch die Vorgangsweise, wie bei der Datenerhebung Stichproben ermittelt bzw. herangezogen werden, kann diesen Bias begünstigen. So kann es beispielsweise passieren, dass eine Stichprobenerhebung nur einen Teil des Datenraums erreicht und diesen nicht vollständig abbildet. Teilweise wird für Repräsentationsbias auch der Begriff Auswahlbias (Sampling Bias oder Selection Bias) verwendet. Gründe für Daten, die Subpopulationen nur unzureichend repräsentieren, sind beispielsweise:

- Es erfolgt eine aktive Auswahl von Untergruppen (kein Zufallsprinzip), dadurch Ausgrenzung anderer.
- Teilnehmer entscheiden selber über ihre eigene Teilnahme („Selbstselektion“), haben aber evtl. völlig andere Merkmale als Nicht-Teilnehmer. In diesem Fall wird auch von Selbstselektionsbias (Self-Selection Bias) geredet.
- Mangels Sorgfalt liegt eine übermäßige Repräsentation von leicht verfügbaren Teilnehmern vor. Hier wird auch von Verfügbarkeitsbias (Availability Bias) geredet.
- Bei zeitabhängigen Daten wird eine Querschnittsanalyse durchgeführt, d.h. anstatt die Datenreihe über die Zeit zu betrachten (Längsschnittstudie), werden lediglich stichprobenartig einzelne Zeitpunkte betrachtet.

Beispiel: Eine Umfrage zum Zustand des öffentlichen Personennahverkehrs (ÖPNV) wird nur in Großstädten durchgeführt. Daten, die aus dieser Umfrage gewonnen werden, bilden den Zustand des ÖPNV in ländlichen Gebieten nicht ab – die Stichprobe ist unzureichend erstellt, sodass der ländliche Raum unterrepräsentiert ist – und somit gibt es einen Repräsentationsbias gegen den ländlichen Raum. Gleichzeitig würde bei einer Umfrage unter den Bürgerinnen und Bürgern Deutschlands ebenfalls die Gefahr bestehen, dass das Ergebnis einen Repräsentationsbias gegen den ländlichen Raum aufweist. Im Jahr 2022 lebten mehr als 70% der deutschen Bevölkerung in Großstadtregionen (Statistisches Bundesamt). Wenn zufällig Menschen für die Umfrage ausgewählt werden, ist die Wahrscheinlichkeit hoch, dass der ländliche Raum weiterhin unterrepräsentiert ist.

Das Beispiel zeigt, wie leicht ein Repräsentationsbias in Daten entstehen kann, sofern dieser bei dem gewählten Ansatz für die Datenerhebung nicht berücksichtigt wird. Zur Vermeidung ist es also wichtig, Teilpopulationen zu identifizieren und - nach technischen Möglichkeiten - auch mit jeweils genügend Stichproben zu berücksichtigen.

A3 Messungsbias (Measurement Bias)

Diese Art von Bias tritt durch Entscheidungen auf, die bei der Auswahl der Messverfahren und der Messung/Berechnung sowie Nutzung/Erfassung von Merkmalen getroffen werden. Außerdem können Messfehler zu Messungsbias führen.

Beispiel: „COMPAS“ ist ein Prognosetool in den USA zur Ermittlung des Rückfallrisikos von Straftätern. Die für die Vorhersage genutzten Merkmale umfassen unter anderem vorherige Verhaftungen, ebenfalls werden Verhaftungen von Freunden und Familienangehörigen für

die Prognosen berücksichtigt, um die „Gefährlichkeit“ bzw. Rückfallgefahr der betreffenden Personen zu messen. Die Relevanz dieser Merkmale für das zu lösende Vorhersageproblem ist höchst umstritten (Angwin, et al., 2016) (Dressel, et al., 2018) (Suresh, et al., 2019). Somit kann die Betrachtung dieser Merkmale zu Messungsbias führen.

A4 Bias durch Weglassen von Variablen (Omitted Variable Bias)

Bias durch Weglassen kann auftreten, wenn wichtige Merkmale bei der Erfassung der Daten nicht berücksichtigt werden oder aus anderen Gründen in den resultierenden Datensätzen fehlen.

Beispiel: Ein Modell soll die Wahrscheinlichkeit berechnen, mit der Patienten bei einer Grippeinfektion einen schweren Krankheitsverlauf entwickeln, berücksichtigt aber nicht das Alter der Patienten. Da dieses Merkmal allerdings positiv mit einem schweren Krankheitsverlauf korreliert, kann das Modell durch das Weglassen des Alters einen Bias entwickeln.

2.2 Bias bei der Entwicklung von KI-Systemen

In diesem Unterkapitel werden Bias-Arten eingeführt, die bei der Konzeption, dem Training und der Evaluation von KI-Modellen auftreten können.

A5 Evaluationsbias (Evaluation Bias)

Die Art der Evaluation eines KI-Modells – sprich, welche Metriken zur Messung des Vorhersageerfolgs genutzt werden - oder Designentscheidungen bei der Evaluationsmethodik können das gelernte Verhalten eines KI-Modells beeinflussen. Auch hat die Wahl der Evaluationsbenchmarks eine Auswirkung auf die Einschätzung des Verhaltens von trainierten KI-Modellen. Wenn durch den Evaluationsprozess eine Ungleichbehandlung von Subpopulationen in das KI-Modell eingeführt wird bzw. nicht erkennbar ist, wird von Evaluationsbias gesprochen.

Beispiel: Eine Bilddatenbank wird für die Bewertung bei Gesichtserkennungssystemen genutzt (beispielsweise die 2014 veröffentlichte Benchmark *Adience*). Die Verwendung dieses Datensatzes erwies sich als unangemessen und unverhältnismäßig, da sich dieser zu fast 80% aus hellhäutigen Gesichtern zusammensetzte. Mit diesem Datensatz evaluierte KI-Modelle hatten einen Bias gegenüber Menschen mit dunkler Hautfarbe (Buolamwini, et al., 2018).

A6 Algorithmischer Bias (Bias through Algorithmic Design)

Die Verwendung von problematischen Daten ist nicht das einzige Einfallstor für Bias. Auch die Auswahl von Trainings- und Evaluationsalgorithmen kann zu Bias in den resultierenden Modellen führen. Da Algorithmen auf unterschiedlichen Annahmen aufbauen, können von den Algorithmen unterschiedliche Muster in den Daten beim Training überbetont oder unterbewertet werden. Hierdurch können die Modelle Bias entwickeln. Zu den algorithmischen Entscheidungen, die Bias einführen können, gehören beispielsweise die Wahl der Optimierungsfunktion, der Einsatz von Regularisierung, die Anwendungsebene von Regressionsmodellen (Datengesamtheit oder Subgruppen) oder die Wahl von Heuristiken. Teilweise findet sich in der Literatur auch der Begriff Algorithmic Bias für diese Bias-Art. Dieser Begriff wird allerdings ebenfalls als Sammelbegriff für Fehler in automatisierten Entscheidungsprozessen genutzt, die zu „unfairen“ Ergebnissen führen.

Beispiel: Die Trainingsprozedur eines KI-Modells benutzt einen Zufallszahlengenerator, beispielsweise für die Auswahl oder Neusortierung der Datenpunkte. Wenn ein unzuverlässiger Algorithmus für die Generierung der Zufallszahlen verwendet wird, können einzelne Datenpunkte oder Subpopulationen innerhalb des Datenraumes unverhältnismäßig

über- oder unterbetont werden. Dies kann zu Bias im Verhalten des trainierten KI-Systems führen, selbst wenn mit großer Sorgfalt auf die Vermeidung von Bias bei der Erstellung der Trainingsdaten geachtet wurde.

A7 Aggregationsbias (Aggregation Bias)

Unter Aggregation versteht man das Zusammenfügen oder Zusammenfassen von Informationen. Beim Aggregationsbias wird letzteres betrachtet. Wenn Modelle zusammenfassende Statistiken des Datenraumes nutzen und nicht die Feinheiten des Datenraumes beachten, kann Bias in den Modellen entstehen. Dies liegt daran, dass zusammenfassende Statistiken die Gefahr haben, wichtige Informationen der einzelnen Datenpunkte oder Subpopulationen zu verlieren bzw. zu verfälschen. Ein berühmtes Beispiel hierfür ist das Simpson-Paradoxon, bei dem Trends innerhalb von Daten-Subpopulationen bei einer Aggregation verschwinden oder sogar umgekehrt werden (Wikipedia).

Beispiel: Eine bekannte Fallstudie, bei der die Aggregation der Daten zu Bias geführt hat, kommt aus dem medizinischen Bereich und untersuchte die Wirkung von Behandlungsmethoden bei Nierensteinen. In der Studie wurden zwei Behandlungsmethoden – der Einfachheit halber nennen wir die Methoden A und B – bei Patienten mit kleinen Nierensteinen und bei Patienten mit großen Nierensteinen getestet. Bei separater Betrachtung der Patientengruppen hat jeweils Methode A einen größeren Behandlungserfolg aufgewiesen, aber durch die Aggregation der Patientengruppen erschien Methode B als die erfolgreichere Behandlung (Charig, et al., 1986). Wird ein KI-Modell auf solch aggregierten Daten trainiert, kann dieser Aggregationsbias auch in das KI-Modell einfließen.

2.3 Bias bei der Nutzung eines KI-Systems

Hier werden Bias-Arten betrachtet, die beim Einsatz eines KI-Systems inklusive der Interaktion mit Nutzenden entstehen können.

A8 Interaktionsbias (User Interaction Bias)

Als Interaktionsbias werden die Formen von Bias bezeichnet, die erst auftreten, wenn Nutzende mit einem KI-System interagieren. Grundsätzlich lassen sich hierbei zwei Arten unterscheiden: Bias kann von Nutzenden auf KI-Modelle übertragen werden oder das Design der Interaktionsschnittstellen begünstigt die Entstehung von Bias.

Damit der erste Fall – die Übertragung von Bias der Nutzenden auf KI-Modelle – auftreten kann, muss gegeben sein, dass die Interaktionen der Nutzenden geloggt werden und anschließend als Trainingsdaten für die nächste Iteration der KI-Modelle verwendet werden. Wenn Nutzende in ihrer Interaktion mit dem KI-Modell verstärkt internalisierte Vorurteile oder Fehlinformationen einbringen, können diese auf das KI-Modell übertragen werden und somit zu einem Bias in der nächsten Iteration des KI-Modells führen. Hierbei ist wichtig zu beachten, dass viele KI-Modelle – z.B. die Sprachmodelle bei Chatbots – dynamisch sind und stetig weitertrainiert werden. Dafür werden auch Informationen aus der Interaktion mit Nutzenden verwendet. Somit behandelt dieser erste Fall eine realistisch auftretende Situation.

Im zweiten Fall – Bias begünstigt durch Designentscheidungen bei der Interaktionsschnittstelle – unterscheiden wir insbesondere zwei Unterarten des Interaktionsbias, nämlich:

- *Präsentationsbias (Presentation Bias)*: Hier wird Bias durch die Art und Weise, wie den Nutzenden Ergebnisse präsentiert werden, begünstigt. Wird nur eine gewisse Auswahl an Inhalten angezeigt, kann somit keine vollständige Darstellung des Datenraumes stattfinden und nicht angezeigte Inhalte erhalten keine Interaktion. Dieser Effekt kann sich weiter verstärken bei dynamisch eingesetzten KI-Modellen, d.h. bei KI-Modellen,

die stetig weitertrainiert werden. Durch die fehlenden Interaktionen mit nicht-angezeigten Inhalten werden diese in den folgenden Iterationen schwächer gewichtet und somit reduziert sich die Wahrscheinlichkeit, dass diese Inhalte angezeigt werden.
- *Auflistungsbias (Ranking Bias)*: Diese Bias-Art ist ähnlicher Natur, allerdings steht hier die Reihenfolge (engl.: ranking) der angezeigten Inhalte im Vordergrund. Ergebnisse, die in einer Auflistung weiter vorne stehen, werden häufiger von Nutzenden angeklickt (Mendler-Dünner, et al., 2024). Anfällig für diese Form des Bias sind Anwendungen, bei denen eine Sortierung in der Anzeige vorgenommen wird. Bekannte Beispiele sind Suchmaschinen, Crowdsourcing-Dienste und Empfehlungssysteme (engl.: recommendation systems), wie sie in Streamingdiensten oder Online-Marktplätzen eingesetzt werden. Für diesen Bias findet sich auch häufig die Bezeichnung Popularitätsbias (Popularity Bias).

A9 Populationsbias (Population Bias)

Der Populationsbias weist Ähnlichkeiten zum Evaluationsbias auf. Auch hier entsteht der Bias durch Abweichungen in den Populationsstrukturen von Daten. Dieser Bias tritt auf, wenn die Strukturen innerhalb der Trainings- und Testdaten (z.B. Demographie, Subpopulationsverteilung, etc.) nicht übereinstimmen mit Strukturen innerhalb von Daten, mit denen das trainierte Modell im Einsatz in Berührung kommt. Vereinfacht gesagt, kann dieser Bias auftreten, wenn ein (möglicherweise) korrekt trainiertes KI-Modell in einem unpassenden Einsatzgebiet verwendet wird.

Beispiel: Eine KI zur Intrusion-Detection wird mithilfe von Aktivitäts- und Netzwerkinformationen von Unternehmen A trainiert. Wird diese KI nun in einem anderen Unternehmen B eingesetzt, kann es sein, dass sich die Populationsstrukturen innerhalb der Aktivitäts- und Netzwerkinformationen von Unternehmen B – d.h. die Definition normaler und maliziöser Datenpunkte – stark vom Unternehmen A unterscheiden. In diesem Fall kann die KI die Intrusion Detektion nicht zuverlässig gewährleisten.

A10 Bias im Einsatz (Deployment Bias)

Als solchen bezeichnet man einen Bias, der durch Zweckentfremdung eines KI-Modells auftreten kann. Wenn ein KI-Modell für eine bestimmte Vorhersageaufgabe trainiert wurde und anschließend für eine andere Vorhersage zweckentfremdet wird, stimmen die Strukturen innerhalb des Datenraumes zwischen Trainings-/Testumgebung und Einsatzumgebung nicht mehr überein. Hierdurch kann ein Bias im Verhalten des KI-Modells entstehen, selbst wenn das KI-Modell bei einem angemessenen Einsatz keinen Bias vorweisen würde.

Beispiel: Ein KI-Modell wird auf Zeitreihendaten trainiert, um Wettervorhersagen zu treffen. In einer Zweckentfremdung wird dieses KI-Modell auf Zeitreihendaten angewendet, um ein automatisches Bewässerungssystem zu steuern. Zwar sind die Eingabedaten gleich, aber die Vorhersageaufgabe – Wetterbedingungen vorhersagen versus Bewässerung optimieren – unterscheiden sich. Dies kann zu Bias im Einsatz beim Modellverhalten führen.

A11 Automationsbias (Automation Bias)

Der Automationsbias beschreibt einen bekannten psychologischen Effekt (Madhavan, et al., 2007). Hierbei entwickeln Menschen ein übermäßiges Vertrauen in automatisierte Prozesse, welches in einem unkritischen Umgang mit automatisierter Entscheidungsfindung resultieren kann. Im Extremfall übernehmen Menschen Ausgaben und Ansichten von Maschinen, selbst wenn diese der eigenen Expertise widersprechen und offensichtlich falsch bzw. unglaubwürdig sind (Jabbour, et al., 2023).

Dieses „blinde“ Vertrauen in die KI und das damit verbundene Einstellen jeden Mitdenkens kann Gefahren mit sich bringen. Problematisch werden Situationen, in denen KI-Systeme als Unterstützung konzipiert werden. Durch Automationsbias kann eine faktische Delegierung der Aufgabe an das KI-System entstehen, wodurch das KI-System – das eigentlich nur unterstützen oder Entscheidungsvorschläge machen soll – faktisch selber Entscheidungen trifft. Der Einfluss des eigentlich verantwortlichen Menschen wird ersatzlos gestrichen. Auch wenn diese Bias-Art nicht direkt das Verhalten von KI-Modellen betrifft, muss Automationsbias immer mitbedacht werden, um problematische bis gefährliche Situationen präventiv zu verhindern.

Beispiel: Ein KI-Modell berechnet Routen für ein Navigationssystem. Automationsbias würde hier bedeuten, dass Autofahrende den Anweisungen des Navigationssystem blind folgen, ohne deren Sinnhaftigkeit zu überprüfen. Die Folgen einer solchen Situation können im schlimmsten Fall fatal sein – das Auto wird beispielsweise in ein Gewässer oder gar den Gegenverkehr gelenkt.

2.4 Weitere Begrifflichkeiten

Wie bereits erwähnt zeichnet sich das Forschungsfeld zum Thema Bias durch eine Vielzahl an verschiedenen Begrifflichkeiten aus. Oftmals handelt es sich bei vorgeschlagenen Bias-Arten um Sonderfälle, Ober- oder Unterkategorien anderer Bias-Arten. Eine vollständige, detaillierte Auflistung aller in der Literatur verwendeten Bezeichnungen ist aus diesem Grund wenig zielführend. Um den Lesenden allerdings den Einstieg zu erleichtern, findet sich im Folgenden eine Auflistung weiterer Begrifflichkeiten, die in einschlägiger Literatur gefunden werden kann.

- Bias in erzeugten Inhalten (Content Production Bias): Dieser Begriff bezieht sich auf Online-Plattformen, z.B. Soziale Netzwerke oder Video-Plattformen, und beschreibt den Bias, der durch von Nutzenden ohne Qualitätskontrollen erzeugte Inhalte entsteht.
- Datenbias: Ein Oberbegriff, mit dem verschiedene Arten aus dem Bereich „Bias bei der Datenerhebung“ zusammengefasst werden.
- Designer- / Entwicklerbias: Ein Oberbegriff, mit dem verschiedene Arten aus dem Bereich „Bias bei der Entwicklung von KI-Systemen“ zusammengefasst werden.
- Emergent Bias: Hiermit wird ein Bias beschrieben, der auftritt, wenn sich die Datenverteilung im Einsatzgebiet eines KI-Systems nach der Einführung des KI-Systems verändert.
- Sozialbias (Social Bias): Dieser Begriff beschreibt den Bias, der durch die Kenntnis vom Verhalten anderer Nutzenden auftritt. Dies umfasst das bekannte psychologische Phänomen, dass Menschen sich dem Verhalten einer Gruppe anpassen und die eigene Einschätzung verändern, um mit der Mehrheit übereinzustimmen (Abels, 2020).
- Verhaltensbias (Behavioral Bias): Dieser Bias tritt auf, wenn wiederkehrende Verhaltensmuster von Menschen deren Entscheidungsfindung beeinflussen und verzerren.
- Verknüpfungsbias (Linking Bias): Dieser Begriff beschreibt den Umstand, dass Aktivität von Nutzenden innerhalb von Sozialen Netzwerken nicht notwendigerweise die Gesinnung der Nutzenden widerspiegeln müssen.
- Zeitbias (Temporal Bias): Dieser Begriff beschreibt Bias, der sich durch Veränderungen von Strukturen (beispielsweise Populationsstrukturen) im Datenraum über die Zeit ergibt.

3 Bias-Detektion

Das Auffinden von Bias in Daten oder KI-Modellen stellt eine große Herausforderung dar, für die es kein einfaches Rezept gibt. Auf der einen Seite sind Trainingsdaten und KI-Modelle nicht uniform und können zwischen verschiedenen Anwendungsgebieten starke Variationen aufweisen. Auf der anderen Seite gibt es sehr vielfältige Bias-Arten und damit verbunden zahlreiche und heterogene Möglichkeiten, mit denen Bias in Daten und KI-Modelle Einzug halten kann. Aus diesen Gründen muss die Detektion von Bias als iterativer und dynamischer Prozess verstanden werden. Verschiedene Werkzeuge aus der statistischen Analyse können für die Bias-Detektion verwendet werden. Neben diesen quantitativen Werkzeugen sind auch qualitative Werkzeuge essentiell für die Bias-Detektion. Dieses Kapitel gibt einen Überblick über die gebräuchlichsten Werkzeuge.

3.1 Detektion von Bias in Daten

Da KI-Modelle statistische Zusammenhänge aus den Trainingsdaten lernen, ist es wichtig, bereits bei den Daten mit der Detektion von Bias zu beginnen. Ziel ist es hierbei, mögliche Bias-Quellen so früh wie möglich zu erkennen und potentiell zu beseitigen, um die Entwicklung eines Bias-freien KI-Modells zu erleichtern.

D1 Qualitative Datenanalyse

Eine qualitative Analyse der Daten besteht aus einer intensiven Beschäftigung mit den Daten selbst. Dies beinhaltet, umfangreiche Kenntnisse über wichtige Eigenschaften der Daten zu erarbeiten. Dazu gehören beispielsweise alle zur Verfügung stehenden Metadaten[5], aber auch Informationen zur Herkunft der Daten, der Datenerhebungsmethodik, möglicherweise durchgeführte Transformationen und viele weitere. Aufgrund der bereits angesprochenen Heterogenität von Daten kann kein allgemein gültiges Verfahren für eine qualitative Datenanalyse bereitgestellt werden. Allerdings sollten **mindestens** folgende Fragen, falls zutreffend, in einem qualitativen Analyseverfahren von Daten beantwortet werden:

- *Wie repräsentativ sind die Daten?* Daten sind immer eine Abstraktion der Realität und die Implikationen dieser Abstraktion müssen eingeschätzt werden.
- *Wann wurden die Daten erhoben? Wie alt ist der neueste Datenpunkt?* Sehr alte Daten können Situationen repräsentieren, die aus moderner Sicht überholt sind.
- *Wie wurden die Daten erhoben?* Verschiedene Methoden haben verschiedene Vor- und Nachteile. Unter Umständen können manche Datenerhebungsmethoden wichtige Merkmale des Datenraumes nicht erfassen. Hierbei sollte auch überprüft werden, wie groß der Anteil von synthetischen/KI-generierten Datenpunkten ist.
- *Wie aufwändig ist die Datenerhebung?* Je aufwändiger die Datenerhebung, desto kostenintensiver und damit unwahrscheinlicher ist eine vollständige Abdeckung des Datenraumes.
- *Falls Daten von Menschen erhoben werden, wissen diese von der Datenerhebung?* Das Wissen um eine Datenerhebung kann bekanntermaßen das Verhalten und somit die Daten beeinflussen.

[5] Als Metadaten werden zusätzliche Informationen über einen Datensatz bezeichnet, die nicht aus den Merkmalen herleitbar sind, aber unter Umständen wichtiges Zusatzwissen beinhalten. Nehmen wir zur Veranschaulichung ein Beispiel aus der Medizin. Ein Datensatz aus der Radiologie beinhaltet CT-Scans des Brustbereichs von Patienten. Die Merkmale sind lediglich die Pixel der Bilder. Metadaten für diesen Datensatz wären beispielsweise das Geschlecht, das Alter oder die Ethnizität der Patienten. Es ist hierbei sehr einfach ersichtlich, dass diese Informationen für eine adäquate Verarbeitung des Datensatzes wichtig sind.

- *Wer hat die Möglichkeit, an der Datenerhebung teilzuhaben?* Je größer die Barrieren sind, um an einer Datenerhebung teilzunehmen, desto wahrscheinlicher wird der Datenraum durch den entstehenden Datensatz ungleichmäßig abgedeckt.
- *Wie ist die (zeitliche, räumliche, etc.) Auflösung der Daten?* Durch eine unpassende Auflösung der Daten können beispielsweise wichtige Muster verdeckt werden.
- *Welche Merkmale der Daten wurden tatsächlich gemessen und welche lediglich ab- oder hergeleitet?* Das Ab- oder Herleiten von Merkmalen aus anderen Messwerten benötigt zusätzliche Annahmen oder Techniken wie Filter und Aggregationen, welche Bias einführen können.
- *Wurden alle relevanten Merkmale der Daten erfasst?* Das Fehlen von wichtigen Merkmalen kann zu Bias in den trainierten KI-Modellen führen (siehe A4: Bias durch Weglassen). Aus diesem Grund sollte eine qualitative Einschätzung vorgenommen werden, ob alle relevanten Merkmale der Daten auch tatsächlich im Datensatz erfasst wurden.
- *Können die Daten Falschinformationen enthalten?* Mit der weiteren Verbreitung von KI-Systemen werden Trainingsdaten ein beliebteres Ziel für Angriffe. Beispielsweise kann das Schutzziel Integrität der Daten angegriffen werden, um gezielt (politische) Desinformationskampagnen zu verstärken. Somit müssen Trainingsdaten auf Falschinformationen überprüft werden.

Wenn eigene Datensätze für das Training von KI-Modellen erstellt werden, sollte eine sogenannte Datenkarte (engl.: datasheet) angefertigt werden (Gebru, et al., 2021). Das Anfertigen einer solchen Datenkarte erfordert ein tiefes Verständnis und grundlegende Kenntnisse der Daten. Außerdem kann damit die Wiederverwendung der Daten erleichtert werden, da potentielle Nutzende des Datensatzes wichtige Informationen über die Daten erhalten.

D2 Quantitative Datenanalyse

Bei der quantitativen Analyse von Daten kommen verschiedene Verfahren aus der Statistik zum Einsatz. Hiermit können statistische Eigenschaften der Daten untersucht werden, die auf die Qualität der Daten und die Wahrscheinlichkeit von Bias schließen lassen. Um einen generellen statistischen Einblick in die Daten zu erhalten, können folgende Schritte verwendet werden:

- Die Schiefe (engl.: skewness) der Datenverteilung sollte untersucht werden. Dies kann mithilfe von Datenvisualisierungen wie beispielsweise Histogrammen oder Dichtekurven realisiert werden. Alternativ gibt es numerische Werte, die für diese Untersuchung genutzt werden können. Dazu gehören der Fischer'sche Schiefekoeffizient, die Pearson'sche Schiefekoeffizienten oder der Quantilskoeffizient der Schiefe. Eine starke Schiefe der Daten stellt generell eine Herausforderung für das Training von KI-Modellen dar, aber liefert außerdem ein Einfallstor für Bias.
- Korrelationen innerhalb der Daten können ebenfalls genutzt werden, um potentiellen Bias festzustellen. Korrelationen zwischen Merkmalen können zusammen mit qualitativem Wissen über die Daten Hinweise zu Bias liefern. Beispielsweise sollten Merkmale, die Korrelationen mit bekannten Bias-Merkmalen (z.B. Ethnizität oder Geschlecht) aufweisen, mit großer Sorgfalt untersucht werden, da diese ebenfalls anfällig für Bias sein könnten. Für Korrelationsuntersuchungen gibt es numerische Werte, die verwendet werden können. Dazu gehört der Pearson'sche Korrelationskoeffizient, der Spearman'sche Rangkorrelationskoeffizient (Spearman'sches Rho) oder der Kendall'sche Rangkorrelationskoeffizient (Kendall'sches Tau).

- Speziell für Zeitreihendaten sollte außerdem eine Autokorrelationanalyse durchgeführt werden. Hierdurch können versteckte Muster identifiziert werden, die Hinweise auf Bias enthalten können. Ein mögliches Verfahren zur Identifikation von Autokorrelation ist der Durbin-Watson-Test[6]. Außerdem ist es möglich, Autokorrelationen mit einem Korrelogramm graphisch darzustellen. Es ist empfehlenswert, diese Analyse zusätzlich für verschiedene Subpopulationen innerhalb der Daten separat durchzuführen, um Subgruppen-spezifische Muster identifizieren zu können, die Bias enthalten könnten.
- Ein weiteres Werkzeug ist außerdem die Untersuchung der Varianz einzelner Merkmale in den Daten. Merkmale mit keiner oder sehr geringer Varianz sollten entfernt werden. Diese Merkmale bieten keinen Mehrwert für ein KI-Modell und können außerdem Bias in ein KI-Modell einführen. Hierbei sollte die Analyse zusätzlich für die verschiedenen Subpopulationen innerhalb der Daten durchgeführt werden. Merkmale mit keiner oder sehr geringer Varianz für einzelne Subpopulationen sollten aus demselben Grund entfernt werden. Zur Untersuchung kann die korrigierte Stichprobenvarianz unter Verwendung der Bessel-Korrektur verwendet werden.
- Um konkrete Subpopulationen innerhalb der Daten zu vergleichen sowie um Probleme und Bias in den Repräsentationen der einzelnen Subpopulationen festzustellen, eignen sich statistische Verfahren wie beispielsweise ANOVA[7] (ANalysis Of VAriance, dt.: Varianzanalyse) oder der Schnelltest nach Tukey[8]. Hierbei muss jedoch in jedem Individualfall sichergestellt werden, dass die Voraussetzungen für die Anwendung der Verfahren gegeben sind und die passende ANOVA-Variante verwendet wird.

3.2 Detektion von Bias in KI-Modellen

Teilweise ist eine detaillierte Untersuchung der Daten nicht möglich. Dies kann beispielsweise der Fall sein, wenn Unternehmen sich KI-Modelle von Drittanbietern einkaufen, um diese in ihre eigenen Produkte und/oder Systeme zu integrieren. In diesem Fall ist lediglich der Zugriff auf das Modell möglich. Aber auch hier sollten KI-Modelle vor dem Einsatz auf potenzielle Bias getestet werden. Hierfür kann auf verschiedene Verfahren zurückgegriffen werden, die helfen sollen, Bias in bereits trainierten KI-Modellen zu identifizieren.

D3 Fairness-Metriken

Eine Möglichkeit, Bias-bedingte Verhaltensauffälligkeiten bei KI-Modellen festzustellen, sind sogenannte Fairness-Metriken. Dies sind statistische Kenngrößen, die das Vorhersageverhalten von KI-Modellen speziell unter Berücksichtigung von sensitiven Attributen quantifizieren. Mit diesen Metriken können verschiedene Arten der Ungleichbehandlung von Subpopulationen durch KI-Modelle identifiziert werden. Im Folgenden werden verschiedene Fairness-Metriken vorgestellt und erklärt. Da zahlreiche Fairness-Metriken vorgeschlagen wurden, werden hier nur die gebräuchlichsten aufgeführt. Der Einfachheit halber wird im Folgenden ein binäres KI-Modell betrachtet, d.h. das KI-Modell

[6] Der Durbin-Watson-Test ermittelt, ob bei einer Regressionsanalyse aufeinanderfolgende Residualwerte miteinander korreliert sind. Damit wird auf das Vorliegen einer Autokorrelation 1. Ordnung getestet.

[7] ANOVA bezeichnet eine Gruppe an statistischen Verfahren, mit denen Subpopulationen innerhalb von Datenräumen verglichen werden können. Hierfür werden Varianzen und weitere Prüfgrößen berechnet, um Gesetzmäßigkeiten innerhalb der untersuchten Daten zu analysieren. Die Gruppe der ANOVA-Verfahren ist sehr groß, so gibt es beispielsweise Versionen für unabhängige Variablen mit festen Effekten oder zufälligen Effekten sowie univariate und multivariate Analysevarianten.

[8] Der Schnelltest nach Tukey ist ein relativ einfach umzusetzendes, nichtparametrisches Verfahren, das zwei unabhängige Stichproben hinsichtlich der Lage ihrer Elemente vergleicht. Zur Berechnung der Testgröße wird die Verschiebung der Elemente beider Stichproben bei einer gemeinsamen Sortierung betrachtet.

ordnet Datenpunkte anhand ihrer Merkmale einer von zwei möglichen Klassen[9] zu. Des Weiteren nehmen wir an, dass es zwei Subpopulationen innerhalb der Daten gibt, d.h. das betrachtete sensitive Attribut kann zwei Werte annehmen. Alle vorgestellten Definitionen können auf komplexere Situationen als den binären Fall erweitert werden. Im Folgenden gilt, dass Y die tatsächliche Klasse eines betrachteten Datenpunkts ist, R ist die vom KI-Modell vorhergesagte Klasse eines Datenpunkts, S ist die vorhergesagte Eintrittswahrscheinlichkeit[10] (engl.: probability score) der vom KI-Modell vorhergesagten Klasse eines Datenpunktes und A bezeichnet das sensitive Attribut der Datenpunkte, d.h. die Subpopulation zu der ein Datenpunkt gehört.

- *Demographische Parität* (engl.: demographic parity oder statistical parity) berücksichtigt die Wahrscheinlichkeit, mit der Datenpunkte aus verschiedenen Subpopulationen der positiven Klasse zugeordnet werden. Um demographische Parität zu erfüllen, muss ein KI-Modell Datenpunkte aus beiden Subpopulationen mit der gleichen Wahrscheinlichkeit der positiven Klasse zuordnen. Mathematisch wird dies folgendermaßen ausgedrückt:
$$P(R = +|A = a) = P(R = +|A = b) \quad \forall a, b \in A$$

- *Prädiktive Parität* (engl.: predictive parity oder outcome test) berücksichtigt die Wahrscheinlichkeit, mit der Datenpunkte, die der positiven Klasse zugeordnet werden, auch zu dieser Klasse gehören. Damit ein KI-Modell prädiktive Parität erfüllt, muss die Wahrscheinlichkeit, mit der Datenpunkte, die der positiven Klasse zugeordnet werden, auch tatsächlich zu der positiven Klasse gehören, für beide Subpopulationen gleich sein. Mathematisch wird dies folgendermaßen ausgedrückt:
$$P(Y = +|R = +, A = a) = P(Y = +|R = +, A = b) \quad \forall a, b \in A$$

- *Ausgeglichene Chancen* (engl.: equalized odds oder disparate mistreatment) berücksichtigt sowohl die positive als auch die negative Klasse. In anderen Worten, diese Metrik sagt aus, ob Datenpunkte, die zur selben Klasse gehören, auch gleich vom KI-Modell behandelt werden; unabhängig von der Subpopulation, zu der die Datenpunkte gehören. Damit ein KI-Modell ausgeglichene Chancen erfüllt, muss die Richtig-positiv-Rate (engl.: True Positive Rate, TPR) und Falsch-positiv-Rate (engl.: False Positive Rate, FPR) zwischen den Subpopulationen gleich sein. Mathematisch wird dies folgendermaßen ausgedrückt:
$$P(R = +|Y = y, A = a) = P(R = +|Y = y, A = b) \quad y \in \{+, -\} \quad \forall a, b \in A$$

- *Gleichheit der Gesamtgenauigkeit* (engl.: overall accuracy equality) nutzt die Vorhersagegenauigkeit des KI-Modells. Damit ein KI-Modell Gleichheit der Gesamtgenauigkeit erfüllt, muss die Vorhersagegenauigkeit des KI-Modells für alle Subpopulationen in den Daten gleich sein. Mathematisch wird dies folgendermaßen ausgedrückt:

[9] In der Fachliteratur wird im binären Fall häufig von einer „positiven“ und einer „negativen“ Klasse gesprochen. Wir verwenden diese Nomenklatur hier ebenfalls. Allerdings ist festzuhalten, dass die Bezeichnungen nicht notwendigerweise mit der Bedeutung der Klassen korrelieren. Bei einem KI-Modell, das Bewerbungen klassifiziert, würde die positive Klasse bedeuten, dass die bewerbende Person eingestellt wird. Bei einem KI-Modell, das als Intrusion-Detection System eingesetzt wird, würde die positive Klasse bedeuten, dass ein Angriff auf das Netzwerk festgestellt wurde. Während im ersten Fall die positive Klasse tatsächlich eine gute Situation beschreibt, ist das im zweiten Fall umgekehrt.

[10] Viele KI-Modelle können nicht nur eine einfache Klassifikation durchführen, sondern geben einen Wahrscheinlichkeitswert aus, der angibt, wie wahrscheinlich der betrachtete Datenpunkt zu verschiedenen Klassen gehört. Diese Wahrscheinlichkeitswerte werden anschließend genutzt, um Datenpunkte bestimmten Klassen zuzuordnen.

$$P(R = Y|A = a) = P(R = Y|A = b) \quad \forall a, b \in A$$

- *Gleichbehandlung* (engl.: treatment equality) berücksichtigt das fehlerhafte Verhalten eines KI-Modells. Dies bedeutet, dass das Verhältnis von Datenpunkten, die fälschlicherweise der positiven Klasse zugeordnet wurden (sogenannte Falsch-Positive oder FP), und Datenpunkten, die fälschlicherweise der negativen Klasse zugeordnet wurden (sogenannte Falsch-Negative oder FN), für beide Subpopulationen gleich sein muss. Mathematisch wird dies folgendermaßen ausgedrückt:
$$\frac{FN_{A=a}}{FP_{A=a}} = \frac{FN_{A=b}}{FP_{A=b}}$$

- *Kalibrierung* (engl.: calibration oder test-fairness) nutzt die vom KI-Modell berechneten Eintrittswahrscheinlichkeiten. Um Kalibrierung zu erfüllen, muss ein KI-Modell garantieren, dass Datenpunkte, für die gleiche Eintrittswahrscheinlichkeiten berechnet werden, auch mit der gleichen Wahrscheinlichkeit zur positiven Klasse gehören - und zwar unabhängig davon, zu welcher Subpopulation sie gehören. Mathematisch kann dies folgendermaßen ausgedrückt werden:
$$P(Y = +|S = s, A = a) = P(Y = +|S = s, A = b) \quad \forall s \in S \quad \forall a, b \in A$$

Es ist wichtig zu beachten, dass ein KI-Modell nicht alle Fairness-Metriken gleichzeitig erfüllen kann. Tatsächlich schließen sich verschiedene Fairness-Metriken teilweise gegenseitig aus (Barocas, et al., 2023). Da jede der hier vorgestellten Fairness-Metriken unterschiedliche Aspekte von Ungleichbehandlung durch KI-Modelle abbildet, liegt es in der Verantwortung von Modell-Entwickelnden und Modell-Anwendenden, die „richtige“ Fairness-Metrik im jeweiligen Anwendungsfall auszuwählen.

D4 Sprachbasierte Detektion

Es gibt verschiedene Ansätze, große Sprachmodelle (Large Language Models, LLMs) auf Bias zu untersuchen. Bei KI-Modellen mit allgemeinem Verwendungszweck (General Purpose Artificial Intelligence, GPAI) sollte jedoch beachtet werden, dass Bias in KI-Modellen oftmals nur in Bezug auf einen konkreten Anwendungsfall sinnvoll definiert werden kann. Dies bedeutet, dass eine Bias-Detektion bei GPAI-Modellen für jeden Anwendungsfall separat erfolgen muss. Verhaltensauffälligkeiten eines LLMs gegenüber Geschlechtsidentitäten können beispielsweise zu einem Bias führen, wenn dieses Modell bei der Bewertung von Bewerbungsunterlagen eingesetzt wird. Sollte dasselbe Modell jedoch in einem Intrusion-Detection System eingesetzt werden, kommt es unter Umständen nicht zu einem Bias im KI-System.

Bei Embedding-basierten Methoden wird die Struktur des Embeddingraums eines LLM untersucht. Spezifisch wird hierbei die Distanz zwischen Zielwörtern (z.B. Berufsbezeichnungen) und sensiblen Attributen (z.B. Geschlecht) überprüft, um mögliche Bias innerhalb des Embeddingraums eines Sprachmodells zu identifizieren. Bekannte Algorithmen aus diesem Bereich sind Word Embedding Association Test (Caliskan, et al., 2017) und der Sentence Encoder Association Test (Dolci, et al., 2023).

Prompt-basierte Methoden verwenden kontrollierte Prompts und eine Ausgabeanalyse, um Bias in LLMs zu detektieren. Bei dem kontrafaktischen Ansatz werden gezielt sensible Wörter (z.B. Geschlechts- oder Ethnizitätsindikatoren) in Prompts verändert und anschließend über eine Auswertung der Veränderungen im Ausgabeverhalten Bias ermittelt (Kusner, et al., 2017) (Boyer, et al., 2023). In eine ähnliche Richtung gehen Template-Ansätze. Hierbei werden vordefinierte Satzvorlagen genutzt, die eine systematische und kontrollierte Bias-Analyse verschiedener Komponenten eines Satzes ermöglichen. Hierbei wird Bias ebenfalls über Veränderungen im LLM-Ausgabeverhalten als Reaktion auf Promptveränderungen detektiert

(Stanczak, et al., 2021) (Dong, et al., 2024). Prompt-basierte Methoden erfordern einen hohen manuellen Aufwand bei der Erstellung der Templates und der Auswertung von Veränderungen im Ausgabeverhalten. Dieser Aufwand kann jedoch durch die Nutzung automatisierter Template-Erstellung verringert werden (Radcliffe, et al., 2024).

4 Bias-Mitigation

Die Entwicklung von technischen Möglichkeiten zur Reduktion oder Verhinderung von Bias und daraus resultierendem verzerrten Verhalten von KI-Modellen ist ein aktives Forschungsfeld. Viele der vorgeschlagenen Lösungen lassen sich dabei in drei Klassen einsortieren – abhängig von dem Zeitpunkt im Entwicklungszyklus eines KI-Produkts, zu dem sie eingesetzt werden. Bei den drei Klassen handelt es sich um Präprozessierungsmethoden, Inprozessierungsmethoden und Postprozessierungsmethoden. Generell sollte bereits bei der Datenerhebung darauf geachtet werden, dass Bias in den entstehenden Daten durch geeignete Verfahren und Designentscheidungen – z.B. bei der Gestaltung von Umfragen, Experimenten oder Aufnahmesystemen – verhindert wird. Wenn dies aufgrund von technischen, rechtlichen oder sozialen Gründen oder dem verbundenen Ressourcenaufwand nur beschränkt möglich ist, können die Methoden in diesem Kapitel helfen, den Bias in den Daten auszugleichen. Mehrere der hier aufgeführten Methoden sind in gängigen Fairness-Toolboxen[11] enthalten, die im Bereich Maschinelles Lernen zum Einsatz kommen.

Zur Veranschaulichung der einzelnen Ansätze wird in diesem Kapitel ein vereinfachtes Beispiel verwendet. Wir betrachten als Beispielaufgabe die Klassifizierung von Datenverkehr. Die Merkmale, auf denen hierbei KI-Modelle trainiert werden, sind quantifizierbare Eigenschaften von Datenpaketen. Diese können beispielsweise Port-Nummer, Protokoll, Type of Service (ToS) Feld bei IPv4 oder Traffic Class Feld bei IPv6 sein. Das binäre Label ist die Kennzeichnung, ob ein Datenverkehr gutartig oder maliziös ist. Bei diesem Beispiel könnte Bias im KI-Modell zur Ungleichbehandlung von Datenverkehr verschiedener Unternehmen führen.

Generell ist es wichtig zu beachten, dass sensitive Merkmale nicht Teil der Merkmale des Datensatzes sein müssen, um Bias in trainierten KI-Modellen zu erzeugen. Es reicht, wenn diese Information impliziert in den Daten abgebildet ist. Tatsächlich führt das Ignorieren von Gruppenzugehörigkeiten, in der Fachsprache „Fairness through Unawareness“ genannt, nicht zur Vermeidung von Bias und Diskriminierung (Barocas, et al., 2023).

4.1 Präprozessierungsmethoden

Unter Präprozessierungsmethoden werden Ansätze zusammengefasst, bei denen die Bias Problematik bereits vor dem Beginn des Trainings behandelt wird. Hierbei handelt es sich also um Methoden, die Veränderungen an den Trainingsdaten vornehmen. Das Ziel dieser Veränderungen besteht darin, dass mit deren Hilfe der Bias in den Trainingsdaten verringert wird und somit KI-Modelle mit möglichst geringem Bias im Verhalten trainiert werden können.

M1 Veränderung der Label

Eine simple Möglichkeit, Bias in den Daten zu begegnen, ist das gezielte Verändern der Label in den Trainingsdaten. Es wird angenommen, dass der Bias sich in der Zuordnung der Datenpunkte zu einzelnen Klassen manifestiert. Im betrachteten Beispiel wird hierbei analysiert, wie das Verhältnis der Label in den Datenverkehren der verschiedenen

[11] Zu den bekannteren Open-Source Fairness-Toolboxen, die häufig auch verschiedene Fairness-Metriken implementieren, gehören:
- FairLearn - https://github.com/fairlearn/fairlearn
- AI Fairness 360 - https://github.com/Trusted-AI/AIF360
- What-If Tool - https://github.com/pair-code/what-if-tool
- FairTest - https://github.com/columbia/fairtest
- LiFT - https://github.com/linkedin/LiFT

Unternehmen ist. Sollten dabei Ungleichheiten festgestellt werden – wenn beispielsweise deutlich mehr Datenverkehr eines bestimmten Unternehmens als maliziös gekennzeichnet ist – können diese angepasst werden. Ein Ansatz dafür wird „Massaging“ genannt. Hierbei werden Datenpunkte identifiziert, welche besonders gut geeignet für eine Labelveränderung sind. Es handelt es sich hierbei häufig um Datenpunkte, die nah an der Entscheidungsgrenze der KI-Modelle liegen (Calders, et al., 2009) (Kamiran, et al., 2012) (Zliobaite, et al., 2011). Ein anderer Ansatz zur Labelveränderung ist, dafür zu sorgen, dass vergleichbare Datenpunkte auch vergleichbare Label bekommen. In anderen Worten, es wird sichergestellt, dass Datenpunkte, die im Datenraum nah beieinander liegen und damit sehr ähnlich sind, auch das gleiche Label zugeordnet bekommen. Hierbei dürfen mögliche sensitive Merkmale im Datensatz natürlich nicht für die Bestimmung der Ähnlichkeit verwendet werden. Für diese Bestimmung können etablierte Clusteringverfahren verwendet werden, beispielsweise K-Means (Luong, et al., 2011).

M2 Veränderung der Merkmale

Anstelle einer Veränderung der Label können auch die Merkmale der Datenpunkte angepasst werden, um gegen Bias in den Daten vorzugehen. Hierbei sollen Unterschiede zwischen den verschiedenen Subpopulationen reduziert werden, die lediglich Chancenungleichheit widerspiegeln und keinen objektiven Wert besitzen. In dem Beispiel dieses Kapitels kann dies auftreten, wenn Datenverkehr eines bestimmten Unternehmens grundsätzlich über einen festen Port geroutet wird. Wenn ein KI-Modell lernt, dass Verkehr über diesen Port eher maliziös ist, wird hierdurch dieses Unternehmen stärker belastet als andere Unternehmen. In den Trainingsdaten kann hier gegengesteuert werden, indem die Portnummern der Datenverkehre verändert werden. Bei dieser Veränderung der Merkmale können entweder Datenpunkte aus allen Subpopulationen angepasst werden (Feldman, et al., 2015) (Johndrow, et al., 2019) oder es werden nur Merkmale bei der benachteiligten Subpopulation verändert (Wang, et al., 2019).

M3 Latent Variable

Die Veränderung von Merkmalen zur Bekämpfung von Bias ist nicht die einzige Möglichkeit. Es können auch neue Merkmale zu den Datenpunkten hinzugefügt werden. Solche neuen Merkmale werden in der Fachsprache häufig *Latent Variables* genannt. Wichtig ist hier, dass die hinzugefügten Merkmale einerseits möglichst frei von Bias und andererseits für die zu lernende Aufgabe relevant sein sollten. Häufig werden *Latent Variables* erstellt, die entweder Informationen über Label oder die Gruppenzugehörigkeit der Datenpunkte enthalten. Diese zusätzliche Information soll dem Erlernen eines Bias beim Modelltraining entgegenwirken. In dem Beispielszenario könnte eine solche *Latent Variable* die Annäherung der Unternehmenszugehörigkeit von Datenverkehren sein, falls diese Information nicht in dem Datensatz enthalten ist. In der Literatur finden sich sowohl Vorschläge zum Ableiten von *Latent Variables* von dem Label (Calders, et al., 2010) (Wei, et al., 2020) (Kehrenberg, et al., 2020) als auch von der Gruppenzugehörigkeit (Oneto, et al., 2019) (Diana, et al., 2022). *Latent Variables* können außerdem helfen, kausale Zusammenhänge in den Daten zu analysieren und diese in die Bias-Mitigation einfließen zu lassen (Kilbertus, et al., 2017).

M4 Sampling

Eine weitere Möglichkeit, Bias vor Beginn des Trainings zu begegnen, ist den Einfluss verschiedener Subpopulationen auf das Training anzupassen. Das Ziel hierbei ist es zu verhindern, dass vorrangig Muster der Mehrheitsgruppe von den KI-Modellen gelernt werden und somit die Vorhersageperformance für Minderheitsgruppen sinkt. In dem Beispielsszenario dieses Kapitels könnten diese Möglichkeiten eingesetzt werden, wenn der Großteil des Datenverkehrs innerhalb eines Netzwerks zu einem einzigen Unternehmen gehört. Somit würde die Mehrheit der Datenpunkte die Strukturen des Datenverkehrs dieses

einen Unternehmens widerspiegeln. Dieses Ungleichgewicht kann zu einer Ungleichbehandlung von Datenverkehr anderer Unternehmen führen. Es gibt zwei grundsätzliche Möglichkeiten, dieses Ungleichgewicht auszugleichen. Auf der einen Seite kann die Zusammensetzung der Trainingsdaten verändert werden, um auszugleichen, dass von bestimmten Subpopulationen deutlich mehr Daten zur Verfügung stehen. Hierbei wird entweder die Anzahl der Datenpunkte der Mehrheitsgruppe reduziert – dieses Verfahren heißt Downsampling (Roh, et al., 2021) (Wang, et al., 2022)– oder die Anzahl der Datenpunkte der Minderheitsgruppen wird erhöht – ein Verfahren, das Upsampling genannt wird. Ein häufig verwendeter Algorithmus für das Upsampling-Verfahren ist SMOTE (Synthetic Minority Over-sampling TEchnique) (Chawla, et al., 2002). Mit SMOTE werden nicht einfach Datenpunkte dupliziert, was zu Problemen beim Training des KI-Modells führen kann, sondern synthetische Datenpunkte erzeugt. Dabei wird sichergestellt, dass die synthetischen Datenpunkte realistisch zur jeweiligen Gruppe gehören könnten. Auf der anderen Seite können Datenpunkte individuell gewichtet werden, um Ungleichheit in den Trainingsdaten auszugleichen. Dieser Ansatz sorgt dafür, dass Datenpunkte mit größeren Gewichten beim Training auch stärker das Verhalten des KI-Modells beeinflussen. Zahlreiche Vorschläge für die Zuordnung der „besten“ Gewichte wurden bereits in der Literatur vorgeschlagen (Krasanakis, et al., 2008) (Jiang, et al., 2020) (Li, et al., 2022).

M5 Lernen von Repräsentationen

Oftmals ist es möglich, Daten auf verschiedene Arten darzustellen. Beispielsweise können zweidimensionale Datenpunkte entweder mit kartesischen Koordinaten beschrieben werden oder mit Polarkoordinaten[12]. Dieses Prinzip bildet die Grundlage für eine weitere Möglichkeit, Bias in den Daten bereits vor Beginn des Trainingsprozesses auszugleichen. Hierbei wird versucht, für die Datenpunkte eine neue Darstellung zu finden. Diese soll möglichst viel des ursprünglichen Informationsgehalts beibehalten – um die Vorhersageperformanz nicht zu beeinträchtigen –, aber gleichzeitig bestehenden Bias in den Daten reduzieren. Die Transformation von der ursprünglichen Darstellung in die neue Darstellung wird dabei durch einen Optimierungsprozess gelernt - ein Vorgang, der in der Fachsprache *Representation Learning* genannt wird. Bei dem Beispiel aus der Bewerbungsauswahl würde dies bedeuten, dass die Merkmale der Datenverkehre in eine andere Darstellung transformiert werden (diese könnte beispielsweise ein hochdimensionaler, reellwertiger Vektor sein). Auf diesen Vektoren wird anschließend ein KI-Modell zur Einordnung der Datenverkehre trainiert. Ein vorgeschlagener Ansatz ist der Algorithmus *Learning Fair Representations* (LFR) (Zemel, et al., 2013). Bei LFR werden sogenannte Prototypen gelernt, mit denen die Datenpunkte dargestellt werden. Um Bias zu mitigieren, benutzt LFR dabei zwei Optimierungsziele: zum einen wird der Informationsgehalt der Prototypen in Bezug auf die Vorhersageaufgabe maximiert und zum anderen sichergestellt, dass die Wahrscheinlichkeit, mit der Datenpunkte aus unterschiedlichen Subpopulationen auf die verschiedenen Prototypen projektiert werden, angeglichen wird. Hierfür wird die Kenngröße statistische Parität verwendet. Das Ziel ist hierbei, dass auf den Prototypen Modelle trainiert werden können, die den Prädiktionstask lösen können und keinen Bias beinhalten. Für den Ansatz *Representation Learning* wurden verschiedene weitere Methoden vorgeschlagen, die unterschiedliche Methoden für die Lösung des Optimierungsproblems nutzen, beispielweise *Adversariales Training* (Feng, et al., 2019) oder *Normalized Flows* (Balunovic, et al., 2021).

[12] Polarkoordinaten sind eine Möglichkeit, Punkte im zweidimensionalen Raum zu beschreiben. Im Unterschied zum kartesischen Koordinatensystem werden Punkte hier nicht durch Orte auf zwei orthogonale Achsen beschrieben, sondern durch den Abstand zum Koordinatenursprung, hier Pol genannt, sowie dem Winkel zu einer vorgegebenen Richtung, hier Polarachse genannt.

4.2 Inprozessierungsmethoden

Unter Inprozessierungsmethoden werden Ansätze zusammengefasst, die Veränderungen in den Trainingsalgorithmen von KI-Modellen vorschlagen. Diese Veränderungen im Training sollen bewirken, dass etwaige Bias in den Trainingsdaten nicht auf das Vorhersageverhalten der trainierten KI-Modelle übertragen werden. Genau wie bei den Präprozessierungsmethoden ist das Ziel, ein KI-Modell mit deutlich reduziertem Bias zu erstellen.

M6 Regularisierung

Ein erster Ansatz, um die Mitigation von Bias während des Trainings zu erreichen, ist Regularisierung. Hierbei werden zu den Funktionen[13], mit denen KI-Modelle trainiert werden, zusätzliche Regularisierungsterme hinzugefügt. Fast jeder Trainingsalgorithmus für KI-Modelle verwendet bereits verschiedene Formen von Regularisierung. Meistens soll dadurch ein Overfitting[14] verhindert werden. Bei der Bias-Mitigation wird stattdessen mithilfe spezieller Regularisierungsterme sichergestellt, dass verschiedene Formen von Bias im trainierten KI-Modell reduziert werden. Hierbei werden häufig sogenannte Fairness Metriken verwendet, um unerwünschtes Verhalten der KI-Modelle auf Subpopulationen innerhalb der Trainingsdaten zu quantifizieren. In der Literatur finden sich verschiedene Vorschläge, wie die genaue mathematische Formulierung solcher Bias-Regularisierungsterme aussehen könnte (Chuang, et al., 2021) (Jiang, et al., 2022). Speziell für Entscheidungsbäume können biasmitigierende Regularisierungsterme verwendet werden, um das Kriterium zum Aufspalten der Baumpfade – auch *Splitting Criteria* genannt – anzupassen (Kamiran, et al., 2010).

M7 Beschränkung

In eine ähnliche Richtung geht die zweite Möglichkeit, Bias während des Trainings zu mitigieren. Dies ist die Verwendung von Beschränkungen im Lösungsraum der Optimierungsfunktion, sogenannte *Constraints*. In einfachen Worten beschrieben handelt es sich bei Constraints um Bedingungen, die von den trainierten KI-Modellen zwingend erfüllt werden müssen. Beispielsweise können damit KI-Modelle mit bestimmten Verhaltensweisen ausgeschlossen werden, auch wenn diese theoretisch ein „besseres“ Ergebnis[15] beim Training erzielen würden. Der Unterschied zur Regularisierung besteht also darin, dass bei dem Regularisierungsansatz das Training von KI-Modellen durch die Regularisierungsterme in eine gewünschte Richtung gelenkt werden soll, während bei der Nutzung von Beschränkungen normal trainiert wird und lediglich der Lösungsraum auf gewünschte Bereiche beschränkt wird. Beispielsweise kann der Lernprozess mithilfe eines sogenannten Metaalgorithmus erweitert werden, der eine Biasbeschränkung[16] als Eingabe bekommt und in das Training einfügt (Celis, et al., 2019). Andere Ansätze wirken mit Beschränkung direkt auf die Entscheidungsgrenze von KI-Modellen (Zafar, et al., 2017) oder lernen statistische Entscheidungsgrundsätze mit Biasbeschränkungen (Kilbertus, et al., 2020).

[13] Hierbei handelt es sich um Optimierungsfunktionen, die gerade beim Training von Neuronalen Netzen üblicherweise Verlustfunktionen (engl. loss functions) genannt werden.

[14] Overfitting bezeichnet den Fall, in dem ein KI-Modell lediglich die Labels der Trainingsdaten auswendig lernt. Das Ergebnis von Overfitting ist ein KI-Modell, das perfekte Vorhersagen auf den Trainingsdaten generiert, aber nicht auf neue Daten generalisieren kann.

[15] Das Wort „besser“ bezieht sich hier auf die Trainingsgenauigkeit.

[16] Dies könnte beispielsweise der maximal erlaubte Bias sein, den ein KI-Modell aufweisen darf. Hierbei wird Bias üblicherweise mithilfe von Fairnessmetriken quantifiziert.

M8 Adversariales Lernen

Eine weitere Möglichkeit der Bias-Mitigation während des Trainings ist die Nutzung von adversarialem Lernen. Dieser Ansatz ist weit verbreitet und findet in verschiedenen Subfeldern der künstlichen Intelligenz Anwendung. Während adversariales Lernen üblicherweise mit dem Versuch verbunden wird, ein KI-Modell robust gegen adversariale Angriffe zu machen, können die gleichen Techniken auch dafür verwendet werden, gegen Bias in KI-Modellen vorzugehen. Das Grundprinzip des adversarialen Lernens ist folgendermaßen: Neben dem gewünschten KI-Modell wird gleichzeitig ein weiteres Modell trainiert, der sogenannte *Adversary* (dt.: Gegenspieler). Während das KI-Modell trainiert wird, um ein Prädiktionsproblem zu lösen, wird der *Adversary* trainiert, um das KI-Modell anzugreifen. Bei der Bias-Mitigation besteht dieser „Angriff" darin, die Ausgabe und/oder interne Repräsentationen des KI-Modells zu nutzen, um sensitive Attribute vorherzusagen und damit potentiellen Bias auszunutzen. Beim adversarialen Lernen werden KI-Modell und Gegenspieler in einem kompetitiven Verfahren optimiert. Dies bedeutet, dass das Ziel bei Bias-Mitigation darin besteht, ein KI-Modell zu trainieren, mit dem die gewünschte Prädiktionsaufgabe gelöst werden kann, ohne einem Gegenspieler zu ermöglichen, sensitive Attribute aus den Informationen des KI-Modells herzuleiten.

Ein bekannter Algorithmus, der diese Maßnahme zur Bias-Mitigation implementiert, ist *Adversarially Reweighted Learning* (ARL) (Lahoti, et al., 2020). ARL nutzt Fehler, die von einem *Adversary* identifiziert werden können, um die Worst-Case-Performanz des KI-Modells auf Subpopulationen der Daten zu verbessern. Ein anderer Ansatz ist *Adversarial Debiasing* (Zhang, et al., 2018). Dieser Ansatz funktioniert wie das Training eines Generative Adversarial Network (GAN, dt.: erzeugendes gegnerisches Netzwerk) mit einer zusätzlichen Eingabe, die angibt, welche Fairnessmetrik zur Mitigation des Bias verwendet werden soll. Es ist auch möglich, nicht zwei komplett unabhängige Modelle zu trainieren, sondern gemeinsame Hidden Layer für KI-Modell und Gegenspieler zu lernen und lediglich unterschiedliche Output Layer zu nutzen (Beutel, et al., 2017).

M9 Ensemble-Lernen

Es ist auch möglich, Erkenntnisse aus dem Ensemble-Lernen zu nutzen, um im Trainingsprozess Bias zu bekämpfen. Ensemble Lernen bezeichnet einen Ansatz des maschinellen Lernens, bei dem nicht ein einzelnes Prädiktionsmodell gelernt wird, sondern ein Modellensemble. Die endgültige Vorhersage bei solchen Ansätzen basiert dann nicht mehr auf der Ausgabe eines einzelnen Modells, sondern auf den verschiedenen Ausgaben des Ensembles. Bei der Bias-Mitigation wird das Modellensemble erstellt, indem für jede Subpopulation in den Trainingsdaten ein gesondertes Modell trainiert wird. Für die endgültige Entscheidung kann dann für jeden Datenpunkt entweder die Ausgabe des spezifischen Modells für die zutreffende Subpopulation genommen werden oder eine kombinierte Vorhersage aus den Ausgaben aller Modelle des Ensembles berechnet werden. In dem fortlaufenden Beispiel in diesem Kapitel würde – falls das sensitive Attribut „Unternehmen" betrachtet wird – jeweils ein Modell für verschiedene Unternehmen trainiert werden. Für die Entscheidung der Einstellung würde dann nicht mehr nur ein Modell verwendet werden, sondern mehrere Modelle.

Ein Problem dieses Ansatzes – vor allem wenn Entscheidungen von den Subpopulationsmodellen getroffen werden und nicht als Ensembleentscheidung – besteht in der Reduktion von Trainingsdaten für jedes einzelne KI-Modell des Ensembles und daraus resultierende Probleme bei der Performanz aufgrund von zu kleinen Datenmengen für das Training. Eine Möglichkeit, dies auszugleichen, besteht darin, Erkenntnisse aus dem Transferlernen zu nutzen (Dwork, et al., 2018) (Ustun, et al., 2019). Wenn für die Entscheidungen die mittlere Ausgabe aller Modelle in dem Ensemble genutzt werden soll,

kann eine sogenannte Pareto-Front[17] von KI-Modellen aufgebaut werden. Diese kann in der Anwendung genutzt werden, um für jeden Anwendungsfall die passenden Biasanforderungen auswählen zu können. Verschiedene Ansätze für den Aufbau einer solchen Pareto-Front finden sich in der Literatur (Liu, et al., 2022) (Blanzeisky, et al., 2022).

M10 Angepasstes Lernen

Eine weitere Kategorie von Methoden zur Bias-Mitigation während des Trainings von KI-Modellen kann mit dem Begriff Angepasstes Lernen zusammengefasst werden. In diese Kategorie fallen Methoden, die entweder bestehende Lernalgorithmen anpassen oder neue Lernalgorithmen entwickeln, jeweils mit dem Ziel, dass trainierte KI-Modelle durch diese neuen Algorithmen weniger Bias in ihrem Verhalten aufweisen. Da das Einschlusskriterium dieser letzten Kategorie sehr weit gefasst ist, finden sich heterogene Ansätze in dieser Gruppe an Bias-Mitigationsmöglichkeiten.

Beispielsweise wurden verschiedene Algorithmen vorgeschlagen, die aktives Lernen verwenden (Noriega-Campero, et al., 2019) (Anahideh, et al., 2022). Aktives Lernen (engl.: *active learning*) bezeichnet einen besonderen Lernansatz im maschinellen Lernen, bei dem Lernalgorithmen bei Bedarf Anfragen an eine Informationsquelle[18] stellen können, um zusätzliche Informationen wie Labels für neue Datenpunkte zu erhalten. Ein weiterer Ansatz, der für Bias-Mitigation verwendet werden kann, ist Lernen mit Ablehnung (engl.: *rejection learning* oder *learning with rejection*) (Madras, et al., 2018) (Lee, et al., 2021). Allgemein formuliert ermöglicht Lernen mit Ablehnung das Trainieren von KI-Modellen, die nur dann eine Entscheidung ausgeben, wenn ein ausreichendes Maß an Zuversicht (engl.: *confidence*) für diese Entscheidung erreicht werden kann. Für die Bias-Mitigation kann diese Zuversicht an den Bias des Modells gekoppelt werden.

Des Weiteren wurden konkrete Lernalgorithmen und -ansätze vorgeschlagen, um Bias von KI-Modellen während dem Training in den Griff zu bekommen. Dazu gehören *Approximate Projection onto Star Sets* (AP-Star) (Martinez, et al., 2020), *Distributionally Robust Optimization* (DRO) (Hashimoto, et al., 2018) und *Multicalibration* (Hébert-Johnson, et al., 2018). Wie bereits angesprochen handelt es sich bei der Kategorie Angepasstes Lernen um eine Sammlung von heterogenen Methoden und Ansätzen. Die hier aufgeführten Beispiele sollen einen Überblick über die Möglichkeiten in dieser Kategorie bieten, aber erheben keinen Anspruch auf Vollständigkeit.

4.3 Postprozessierungsmethoden

Unter Postprozessierungsmethoden werden Ansätze zusammengefasst, die erst nach dem Training eines KI-Modells zum Einsatz kommen. Erst nachdem ein fertiges KI-Modell existiert, wird versucht, mit nachgeschalteten Verfahren einen potentiellen Bias in dem KI-Modell auszugleichen.

M11 Korrektur der Eingabe

Die erste Möglichkeit zur nachgeschalteten Bias-Mitigation ist die Veränderung der Eingabedaten. Diese Möglichkeit wurde bereits bei den Präprozessierungmethoden

[17] Die Pareto-Front (auch Pareto-Menge) ist die Menge aller Pareto-Optima. Ein Pareto-Optima ist ein (bestmöglicher) Zustand einer Lösung, wenn verschiedene Zieleigenschaften optimiert werden. In dieser Publikation werden für die KI-Modelle als Zieleigenschaft Performanz und Bias betrachtet. Also handelt es sich bei einem KI-Modell um ein Pareto-Optima, wenn keine der beiden Eigenschaften verbessert werden kann, ohne die andere zu verschlechtern. Alle KI-Modelle, die diese Bedingung erfüllen, bilden zusammen die Pareto-Front.

[18] Diese Informationsquellen können beispielsweise menschliche Nutzende oder kurierte Datenbanken sein.

beschrieben. Während dort jedoch Veränderungen an den Trainingsdaten vorgenommen werden, sind es hier die Eingabedaten während der Test- und Einsatzphase, die verändert werden, um vorhandenen Bias der KI-Modelle auszugleichen. Bei dem Beispiel mit der Einordnung von Datenverkehr würde dies bedeuten, dass der Einsatz eines KI-Modells mit bekannten Biasproblemen ausgeglichen werden soll, indem neue Datenverkehre systematisch verändert werden, bevor diese an das KI-Modell gegeben werden. Theoretisch können verschiedene Methoden aus der Präprozessierungsphase auf den Einsatz in der Postprozessierungsphase übertragen werden; konkret gibt es Vorschläge für perturbationsbasierte Lösungen (Adler, et al., 2018) (Li, et al., 2022). Speziell für generative KI-Modelle gehört in diese Kategorie auch der Vorgang des Systemprompting. Hierbei werden neue Eingaben beim Einsatz eines generativen KI-Modells mit Systemprompts ergänzt, um das Verhalten des KI-Modells zu verändern.

M12 Korrektur der KI-Modelle

Abgesehen von der Veränderung von Eingabedaten können auch trainierte KI-Modelle verändert werden, um Bias im Verhalten der KI-Modelle zu bekämpfen. In dem Anwendungsbeispiel dieses Kapitels würde dies bedeuten, dass ein KI-Modell mit Ungleichbehandlung der Datenverkehre verschiedener Unternehmen trainiert wurde. Dieses Problem soll nun nicht durch die Entwicklung eines neuen KI-Modells ohne das problematische Verhalten gelöst werden, sondern indem nachgeschaltet Verfahren zum Ausgleich des Verhaltens etabliert werden. Es gibt vorgeschlagene Methoden, die spezifisch für bestimmte Modelltypen angewendet werden können. Dazu gehören Methoden zur Berechnung von biasmitigierenden Veränderungen für Neuronale Netze (Du, et al., 2021) (Savani, et al., 2020) (Marcinkevics, et al., 2022), Entscheidungsbäume (Kamiran, et al., 2010) (Kanamori, et al., 2021) oder verschiedene Regressionsverfahren (Jiang, et al., 2020) (Chzhen, et al., 2020). Alternativ gibt es auch modell-agnostische Methoden zur Veränderung von KI-Modellen. Beispielsweise wurde ein Verfahren vorgeschlagen, das Ähnlichkeiten zu *Boosting*[19] aufweist und in einer iterativen Postprozessierung ein bereits trainiertes KI-Modellen von Bias befreien soll (Kim, et al., 2019). Andere Arbeiten gehen ebenfalls in die Richtung, die Mitigation von Bias in einem trainierten KI-Modell über ein Optimierungsproblem in der Postprozessierungsphase zu lösen (Hardt, et al., 2016) (Woodworth, et al., 2017). Es ist ebenfalls möglich, ein trainiertes KI-Modell nachträglich aufzuteilen, d.h. die Anwendung des KI-Modells auf verschiedene Subpopulationen der Daten zu separieren, und anschließend diese Aufteilung zu nutzen, um das Modell zu kalibrieren. Mithilfe einer solchen Kalibrierung kann Bias im Verhalten von KI-Modellen nachträglich reduziert werden (Pleiss, et al., 2017) (Noriega-Campero, et al., 2019).

M13 Korrektur der Ausgabe

Eine weitere Möglichkeit, dem Bias von bereits trainierten KI-Modellen entgegenzuwirken, ist die Veränderung der Ausgabe dieser Modelle. In Gegensatz zu den ersten beiden Arten von Postprozessierungsmethoden werden hierbei also weder die Eingabedaten noch die Modelle verändert, ein Eingriff findet lediglich direkt bei den Ausgaben statt. In dem Beispiel aus der Einordnung von Datenverkehr würde dies bedeuten, dass einzelne Entscheidungen des KI-Modells gezielt geändert werden, um etwaigen Bias im Modell auszugleichen. Verschiedene Vorgehensweisen wurden für diese gezielte Veränderung von Entscheidungen vorgeschlagen. Beispielsweise wird versucht, Datenpunkte zu identifizieren, bei denen die Entscheidung des

[19] Bei *Boosting* handelt es sich um einen bekannten und stark verbreiteten Algorithmus aus dem Ensemble Lernen. Die grundlegende Funktionsweise von *Boosting* besteht darin, aus einer Menge von schwachen Klassifikationen einen stärkeren Klassifikator abzuleiten. Die Kategorien „schwach“ und „stark“ beziehen sich hierbei auf die Performanz der Klassifikatoren auf dem betrachteten Vorhersageproblem.

KI-Modells mit hoher Wahrscheinlichkeit durch Bias beeinflusst wurde, und entsprechende Änderungen an der Ausgabe des KI-Modells werden vorgenommen (Kamiran, et al., 2012) (Lohia, et al., 2019) (Fish, et al., 2016). Andere Ansätze versuchen generelle Regeln zur Modifikation der Ausgabe abzuleiten, um den Bias des KI-Modells auszugleichen (Pedreschi, et al., 2009) (Menon, et al., 2018) (Chiappa, 2019).

5 Bias und Cybersicherheit

Die Ungleichbehandlungen von Subpopulationen durch das Verhalten von KI-Systemen mit Bias können in allen Einsatzgebieten von KI-Systemen zu Problemen führen. Unter anderem entstehen durch Bias auch negative Implikationen für die Sicherheit von und durch KI-Systeme. Im nachfolgenden Kapitel werden Wechselwirkungen zwischen Bias und Cybersicherheit anhand der drei wichtigsten Schutzziele der IT-Sicherheit aufgezeigt. Diese drei Schutzziele werden in der CIA-Triade zusammengefasst: Confidentiality (dt.: Vertraulichkeit), Integrity (dt.: Integrität) und Availability (dt.: Verfügbarkeit) (BSI).

Die in diesem Kapitel aufgezeigten Sicherheitsrisiken verdeutlichen, dass Bias in KI-Systemen nicht ausschließlich aus Sicht von Chancengleichheit und Antidiskriminierungsmaßnahmen relevant ist. Vielmehr sollten KI-Verantwortliche auch für die Sicherheit der KI-Systeme die in diesem Dokument aufgezeigten Maßnahmen zur Detektion und Bekämpfung von Bias umsetzen.

5.1 Auswirkung auf das Schutzziel Vertraulichkeit

Wenn KI-Systeme auf sensiblen Daten trainiert werden, muss die Vertraulichkeit der Daten sichergestellt werden. Aber auch das KI-System selber kann unter Umständen als sensible Information gelten, dessen Vertraulichkeit geschützt werden muss. Bias in KI-Systemen kann jedoch sowohl die Vertraulichkeit der Daten als auch die des KI-Systems kompromittieren.

In Bezug auf Trainingsdaten gibt es eine Wechselwirkung zwischen Bias und Angriffen, die Trainingsdaten aus einem Modell ausleiten. Es konnte gezeigt werden, dass die Anfälligkeit für Membership Inference Attacken nicht gleichmäßig im Datenraum verteilt ist. Unbalancierte Subpopulationen können zu einer stärkeren Gefährdung von einzelnen Gruppen innerhalb der Trainingsdaten führen (Kulynych, et al., 2022) (Zhang, et al., 2020) (BSI, 2022). Somit muss für eine adäquate Risikoabschätzung in Bezug auf Datenvertraulichkeit die Auswirkung von Bias auf Membership Inference Attacken berücksichtigt werden. Im schlimmsten Fall kann ein Angreifer den Bias in einem KI-Modell ausnutzen, um sensible Daten aus dem Modell auszuleiten. Auch für Modell Inversion Attacken können Bias-bedingte Verhaltensauffälligkeiten ein zusätzliches Einfallstor für Angreifer darstellen (Fang, et al., 2024). Im Unterschied zu Membership Inference Attacken – bei denen konkrete Trainingsdatenpunkte ausgeleitet werden – werden bei Model Inversion Attacken allgemein sensible Informationen über verwendete Datensätze aus einem KI-Modell ausgeleitet. Auch die Vertraulichkeit von KI-Systemen könnte durch Bias gefährdet werden. Viele sogenannte Model Extraction Attacken basieren auf Wissensdestillation (eng.: *knowledge distillation*) - einer Technik, bei der ein sogenanntes Schülermodell trainiert wird, um das Verhalten eines Lehrermodells nachzuahmen. Es gibt Hinweise, dass dieser Prozess durch die Exploitation von Bias verbessert werden kann (Lu, et al., 2025). Maßnahmen zur Bias-Mitigation aus Kapitel 4 können verwendet werden, um zu verhindern, dass Bias zum Stehlen von KI-Modellen ausgenutzt wird.

5.2 Auswirkung auf das Schutzziel Integrität

Für Hochrisiko-KI muss die korrekte Funktionsweise garantiert werden. Andernfalls ist die Integrität der Anwendung nicht gewährleistet. Hierbei entstehen, in den Worten der KI-VO, Risiken für die Gesundheit Sicherheit und die Grundrechte. Aber auch in weniger kritischen Anwendungsfällen kann die Beschädigung der Integrität eines KI-Systems zu Sicherheitsvorfällen führen, wenn hierdurch beispielsweise neue Angriffsvektoren für Cyberkriminelle entstehen. Die Ungleichbehandlung von Subpopulationen im Datenraum durch Bias kann die Integrität von KI-Systemen gefährden.

Das Beispiel in der Einleitung zeigt auf, wie Zugangskontrollsoftware durch Bias an Zuverlässigkeit verliert. Die Problematik, dass Repräsentationsprobleme in den Trainingsdaten zu starken Leistungsschwankungen von KI-Systemen für die Gesichtserkennung führen, ist bekannt (Buolamwini, et al., 2018). Doch nicht nur Repräsentationsbias ist eine ernstzunehmende Gefahr für IT-Sicherheitsanwendungen. Historischer Bias kann ebenfalls zu einem Leistungsabfall von KI-Systemen in der Cybersicherheit führen. Dies kann an dem Beispiel der Detektion von maliziösen Datenverkehren deutlich gemacht werden. Es gibt verschiedene Länder, in denen bereits seit mehreren Jahren ein erhöhtes Aufkommen an Cyberkriminalität bekannt ist (Bruce, et al., 2024). Wenn KI-Systeme zur Detektion von maliziösem Datenverkehr auf Trainingsdaten beruhen, die geografische Lokalisation der Angriffe explizit oder implizit – durch Proxy-Variablen – beinhalten, kann eine Verzerrung zu bestimmten Weltregionen stattfinden. Durch immer leistungsfähigere und kostengünstigere Möglichkeiten, die eigene geographische Lokalisation mithilfe von Virtual Private Networks (VPN) zu verschleiern, können solche Verzerrungen die Detektionsleistung der KI-Systeme reduzieren. Das Ergebnis kann sowohl eine erhöhte Falsch-Positiv-Rate sein (d.h. gutartiger Datenverkehr aus bestimmten Weltregionen wird als maliziös eingestuft) als auch eine erhöhte Falsch-Negativ-Rate (d.h. maliziöser Datenverkehr wird nicht erkannt, da er aus einer „sicheren“ Weltregion kommt). Gerade bei verschlüsseltem Datenverkehr muss für die Klassifikation auf statistische Charakteristika zurückgegriffen werden, wodurch die Gefahr für historischen Bias steigt (Ji, et al., 2024).

Verschiedene Gruppen von Angriffsvektoren auf KI-Systeme nutzen durch Bias induziertes Verhalten gezielt aus, um bestimmte Muster in den Ausgaben der KI-Systeme zu erzwingen. Zu den bekanntesten Angriffsvektoren zählen hierbei die Poisoning Attacken und Evasion Attacken (BSI, 2025). Das Ziel eines solchen Angriffs ist es, einen gezielten Bias des KI-Systems gegenüber vom Angreifer definierten Auslösern (engl.: Triggers) zu etablieren. Sobald ein Auslöser vom KI-System verarbeitet wird, zeigen sich festgelegte Muster im Ausgabeverhalten des Systems. Während Poisoning Attacken traditionellerweise darauf abzielen, vordefinierte Muster im Ausgabeverhalten von KI-Systemen zu erzeugen[20], zielen Evasion Attacken häufig darauf ab, Schutzmechanismen zu umgehen oder generell unerwünschte Verhaltensweisen zu provozieren. Poisoning Attacken können genutzt werden, um sowohl generative KI als auch klassische KI anzugreifen (BSI, 2025) (Yerlikaya, et al., 2022). Evasion Attacken werden herkömmlicherweise auf generative KI angewendet.

5.3 Auswirkung auf das Schutzziel Verfügbarkeit

Die Wechselwirkung von Bias mit dem Schutzziel Verfügbarkeit ist aktuell nicht abschließend einschätzbar. Hierfür fehlen belastbare wissenschaftliche oder angewandte Studien, die diese Wechselwirkung untersuchen. Eine potentielle Wechselwirkung entsteht durch die bereits erwähnten Poisoning Attacken. Wenn hierdurch die Integrität eines KI-Systems zu einem ausreichenden Maß gestört wird, ist dieses KI-System nicht mehr einsetzbar und steht somit nicht mehr zur Verfügung. Dynamische KI-Systeme können in einer koordinierten Anstrengung angegriffen werden, um vordefinierte Verhaltensweisen oder reduzierte Performanz zu erzwingen (Vincent, et al., 2021) (Hardt, et al., 2023). Während diese Möglichkeit in der Forschung bisher hauptsächlich im digital-aktivistischen Bereich untersucht wird, ist es theoretisch vorstellbar, dass böswillige Akteure diese Erkenntnisse ausnutzen, um die Verfügbarkeit und Integrität von KI-Systemen anzugreifen.

[20] Im Kontext von LLMs kann es sich dabei beispielsweise um einen konkreten Text handeln, der in die Ausgabe eingebaut wird. Bei einem Klassifikator kann es sich beispielsweise um eine vordefinierte Klasse handeln, die als Reaktion auf den Auslöser ausgegeben wird.

6 Fazit

Bias kann in erheblichen Maßen das Verhalten und die Funktion von KI-Systemen beeinflussen - häufig zum schlechteren. Hierdurch können Schäden für Bürgerinnen und Bürger entstehen, wenn diskriminierende oder fehlerhafte Entscheidungen zu Benachteiligung führen. Außerdem müssen Prozesse in der öffentlichen Verwaltung eine Gleichbehandlung garantieren und Systeme, die im öffentlichen Sektor zum Einsatz kommen, müssen zuverlässig und sicher arbeiten. Solche Garantien kann ein KI-System mit Bias im Verhalten nicht erfüllen. Aber auch für Unternehmen können beträchtliche Schäden entstehen, wenn beispielsweise durch Bias gefärbte Entscheidungen von eingesetzten KI-Systemen zu Störungen im Betriebsablauf oder schadensersatzpflichtigen Fehlern bei angebotenen Produkten führen. Außerdem kann die Cybersicherheit von KI-Systemen durch Bias gefährdet sein. Das macht die Detektion und Mitigation von Bias in KI-Systemen besonders relevant für den Schutz der Interessen von Bürgerinnen, Bürgern und Unternehmen, aber auch für den Einsatz im öffentlichen Sektor und damit für eine konsequente und sichere Digitalisierung.

Es gibt eine Vielzahl an unterschiedlichen Bias-Arten. Diese lassen sich verschiedenen Phasen des Lebenszyklus eines KI-Systems zuordnen. Die erste Phase, in der Bias-Arten auftreten können, ist die Datenerhebung. Die hier auftretenden Bias-Arten führen zu Verzerrungen in Datensätzen, welche sich durch Training auf KI-Systeme übertragen können. Auch bei der Modellentwicklung kann Bias in ein KI-System gelangen. Durch Designentscheidungen bei der Konzeption und dem Training von KI-Modellen kann Bias im Verhalten der resultierenden KI-Modelle entstehen. Abschließend werden Bias-Arten, die beim Einsatz eines KI-Systems auftreten, vorgestellt. Hier kann beispielsweise durch Veränderungen im Datenraum der Anwendungsfälle oder bei der Interaktion mit Nutzenden Bias entstehen. Die Diversität der Bias-Arten bedeutet auch eine Vielzahl von potentiellen Einfallstoren für Bias in KI-Systeme. Aus diesem Grund ist Kenntnis der Bias-Arten wichtig, um diese Einfallstore identifizieren zu können.

Da KI-Modelle Bias aus den Daten übernehmen können, muss die Detektion von Bias auch bei den Daten beginnen. Hierbei wird idealerweise eine Kombination aus qualitativen und quantitativen Analyseverfahren kombiniert, um ein tieferes Verständnis von den Daten zu erhalten und potentielle Einfallstore für Bias identifizieren zu können. Aber auch für bereits trainierte KI-Modelle gibt es Möglichkeiten, diese auf Bias zu untersuchen. Ein wichtiges Werkzeug dafür sind statistische Untersuchungen des Vorhersageverhaltens über Fairness-Metriken. In jedem Fall ist die Untersuchung von Daten und KI-Modellen auf Bias ein elementarer Schritt in der Entwicklung und dem Einsatz von KI-Systemen und sollte von den betroffenen Stakeholdern von KI-Systemen umgesetzt werden.

Um Bias in KI-Modellen zu mitigieren, gibt es verschiedene Möglichkeiten, die sich im Ansatz und dem Zeitpunkt des Einsatzes unterscheiden. Generell können die Maßnahmen zur Bias-Mitigation an drei verschiedenen Zeitpunkten eingesetzt werden: bevor das Training des KI-Modells gestartet wird (Präprozessierungsmethoden), während des Trainings eines KI-Modells (Inprozessierungsmethoden) oder nachdem ein KI-Modell trainiert wurde (Postprozessierungsmethoden). In diesem Manuskript wurden verschiedene Ansätze vorgestellt, die verschiedene Vor- und Nachteile mit sich bringen. Im Einzelfall liegt es in der Verantwortung der Entwickelnden, angemessene Methoden zur Bias-Mitigation anzuwenden. Generell kann keine der Methoden die Mitigation von Bias garantieren. Aus diesem Grund müssen Entwickelnde den Erfolg der ergriffenen Mitigationsmaßnahmen unvoreingenommen evaluieren und im Zweifel verschiedene Methoden ausprobieren, um zufriedenstellende Ergebnisse zu erzielen.

Literaturverzeichnis

Abels, Heinz. 2020. Außenleitung - die Orientierung an den vielen Anderen. *Soziale Interaktion.* 2020.

Adler, Philip, et al. 2018. Auditing black-box models for indirect influence. *Knowledge and Information Systems.* 2018, Bd. 54.

Anahideh, Hadis, Asudeh, Abolfazl und Thirumuruganathan, Saravanan. 2022. Fair active learning. *Expert Systems with Applications.* 2022.

Angwin, Julia, et al. 2016. Machine Bias. *ProPublica.* [Online] 23. Mai 2016. [Zitat vom: 12. Juni 2024.] https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Balunovic, Mislav, Ruoss, Anian und Vechev, Martin. 2021. Fair normalizing flows. *arXiv preprint arXiv:2106.05937.* 2021.

Barocas, Solon, Hardt, Moritz und Narayanan, Arvind. 2023. *Fairness and machine learning: Limitations and opportunities.* Cambridge : MIT Press, 2023. ISBN: 9780262048613.

Beutel, Alex, et al. 2017. Data decisions and theoretical implications when adversarially learning fair representations. *arXiv preprint arXiv:1707.00075.* 2017.

Blanzeisky, William und Cunningham, Pádraig. 2022. Using Pareto simulated annealing to address algorithmic bias in machine learning. *The Knowledge Engineering Review.* 2022, 37.

Boyer, Christopher B, Dahabreh, Issa J und Steingrimsson, Jon A. 2023. Assessing model performance for counterfactual predictions. *arXiv preprint arXiv:2308.13026.* 2023.

Bruce, Miranda, et al. 2024. Mapping the global geography of cybercrime with the World Cybercrime Index. *Plos one.* 2024, Bd. 19.

BSI. 2025. *Generative AI Models - Opportunities and Risks for Industry and Authorities.* Bonn : Bundesamt für Sicherheit in der Informationstechnik, 2025.

—. Lerneinheit 4.1: Grundlegende Definitionen. [Online] Bundesamt für Sicherheit in der Informationstechnik. [Zitat vom: 04. 06 2025.] https://www.bsi.bund.de/DE/Themen/Unternehmen-und-Organisationen/Standards-und-Zertifizierung/IT-Grundschutz/Zertifizierte-Informationssicherheit/IT-Grundschutzschulung/Online-Kurs-IT-Grundschutz/Lektion_4_Schutzbedarfsfeststellung/4_01_Definitionen.html.

—. 2022. *Security of AI-Systems: Fundamentals - Provision or use of external data or trained models.* Bonn : Bundesamt für Sicherheit in der Informationstechnik, 2022.

Buolamwini, Joy und Gebru, Timnit. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. *Conference on fairness, accountability and transparency.* 2018.

Calders, Toon und Verwer, Sicco. 2010. Three naive Bayes approaches for discrimination-free classification. *Data mining and knowledge discovery.* 21, 2010, Bd. 2.

Calders, Toon, Kamiran, Faisal und Pechenizkiy, Mykola. 2009. Building classifiers with independency constraints. *IEEE international conference on data mining workshops.* 2009.

Caliskan, Aylin, Bryson, Joanna J und Narayanan, Arvind. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science.* 2017, Bd. 356.

Celis, Elisa L, et al. 2019. Classification with fairness constraints: A meta-algorithm with provable guarantees. *Proceedings of the conference on fairness, accountability, and transparency.* 2019.

Charig, Clive R, et al. 1986. Comparison of treatment of renal calculi by open surgery, percutaneous nephrolithotomy, and extracorporeal shockwave lithotripsy. *British Medical Journal (Clinical Research Edition)* . 1986, Bd. 292.

Chawla, Nitesh V, et al. 2002. SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research.* 16, 2002.

Chiappa, Silvia. 2019. Path-specific counterfactual fairness. *Proceedings of the AAAI conference on artificial intelligence.* 2019.

Chuang, Ching-Yao und Mroueh, Youssef. 2021. Fair mixup: Fairness via interpolation. *International Conference on Learning Representations.* 2021.

Chzhen, Evgenii, et al. 2020. Fair regression with wasserstein barycenters. *Advances in Neural Information Processing Systems.* 2020.

Dastin, Jeffrey. 2022. Amazon scraps secret AI recruiting tool that showed bias against women. *Ethics of data and analytics.* 2022.

Deutscher Ethikrat. 2023. *Mensch und Maschine - Herausforderungen durch Künstliche Intelligenz.* Berlin : Deutscher Ethikrat, 2023.

Diana, Emily, et al. 2022. Multiaccurate proxies for downstream fairness. *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency.* 2022.

Dolci, Tommaso, Azzalini, Fabio und Tanelli, Mara. 2023. Improving gender-related fairness in sentence encoders: A semantics-based approach. *Data Science and Engineering.* 2023, Bd. 8.

Dong, Xiangjue, et al. 2024. Disclosure and mitigation of gender bias in llms. *arXiv preprint arXiv:2402.11190.* 2024.

Dressel, Julia und Farid, Hany. 2018. The accuracy, fairness, and limits of predicting recidivism. *Science advances.* 2018.

Du, Mengnan, et al. 2021. Fairness via representation neutralization. *Advances in Neural Information Processing Systems.* 2021.

Dwork, Cynthia, et al. 2018. Decoupled classifiers for group-fair and efficient machine learning. *Conference on fairness, accountability and transparency.* 2018.

Fang, Hao, et al. 2024. Privacy leakage on dnns: A survey of model inversion attacks and defenses. *arXiv preprint arXiv:2402.04013.* 2024.

Feldman, Michael, et al. 2015. Certifying and removing disparate impact. *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining.* 2015.

Feng, Rui, et al. 2019. Learning fair representations via an adversarial framework. *arXiv preprint arXiv:1904.13341.* 2019.

Fish, Benjamin, Kun, Jeremy und Lelkes, Ádám D. 2016. A confidence-based approach for balancing fairness and accuracy. *Proceedings of the 2016 SIAM international conference on data mining.* 2016.

Gebru, Timnit, et al. 2021. Datasheets for datasets. *Communications of the ACM.* 2021, Bd. 64.

Hardt, Moritz und Mendler-Dünner, Celestine. 2023. Performative prediction: Past and future. *arXiv preprint arXiv:2310.16608.* 2023.

Hardt, Moritz, Price, Eric und Srebro, Nati. 2016. Equality of opportunity in supervised learning. *Advances in neural information processing systems.* 2016.

Hashimoto, Tatsunori, et al. 2018. Fairness without demographics in repeated loss minimization. *International Conference on Machine Learning.* 2018.

Hébert-Johnson, Ursula, et al. 2018. Multicalibration: Calibration for the (computationally-identifiable) masses. *International Conference on Machine Learning.* 2018.

ISO/IEC TR 24027. 2021. Bias in AI systems and AI aided decision making. Geneva : ISO/IEC, 2021.

Jabbour, Sarah, et al. 2023. Measuring the impact of AI in the diagnosis of hospitalized patients: a randomized clinical vignette survey study. *Jama.* 2023, Bd. 330.

Ji, Il Hwan, et al. 2024. Artificial intelligence-based anomaly detection technology over encrypted traffic: A systematic literature review. *Sensors.* 2024, Bd. 24.

Jiang, Heinrich und Nachum, Ofir. 2020. Identifying and correcting label bias in machine learning. *International conference on artificial intelligence and statistics.* 2020.

Jiang, Ray, et al. 2020. Wasserstein fair classification. *Uncertainty in artificial intelligence.* 2020.

Jiang, Zhimeng, et al. 2022. Generalized demographic parity for group fairness. *International Conference on Learning Representations.* 2022.

Johndrow, James E und Lum, Kristian. 2019. An algorithm for removing sensitive information. *The Annals of Applied Statistics.* 13, 2019, Bd. 1.

Kamiran, Faisal und Calders, Toon. 2012. Data preprocessing techniques for classification without discrimination. *Knowledge and information systems.* 33, 2012, Bd. 1.

Kamiran, Faisal, Calders, Toon und Pechenizkiy, Mykola. 2010. Discrimination aware decision tree learning. *IEEE international conference on data mining.* 2010.

Kamiran, Faisal, Karim, Asim und Zhang, Xiangliang. 2012. Decision theory for discrimination-aware classification. *IEEE 12th international conference on data mining.* 2012.

Kanamori, Kentaro und Arimura, Hiroki. 2021. Fairness-aware decision tree editing based on mixed-integer linear optimization. *Transactions of the Japanese Society for Artificial Intelligence.* 2021, 36.

Kehrenberg, Thomas, Chen, Zexun und Quadrianto, Novi. 2020. Tuning fairness by balancing target labels. *Frontiers in artificial intelligence.* 3, 2020.

Kilbertus, Niki, et al. 2017. Avoiding discrimination through causal reasoning. *Advances in neural information processing systems.* 30, 2017.

Kilbertus, Niki, et al. 2020. Fair decisions despite imperfect predictions. *International Conference on Artificial Intelligence and Statistics.* 2020.

Kim, Michael P, Ghorbani, Amirata und Zou, James. 2019. Multiaccuracy: Black-box post-processing for fairness in classification. *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society.* 2019.

Krasanakis, Emmanouil, et al. 2008. Adaptive sensitive reweighting to mitigate bias in fairness-aware classification. *Proceedings of the 2018 world wide web conference.* 2008.

Kulynych, Bogdan, et al. 2022. Disparate vulnerability: On the unfairness of privacy attacks against machine learning. *Proceedings of 22nd Privacy Enhancing Technologies Symposium.* 2022.

Kusner, Matt J, et al. 2017. Counterfactual fairness. *Advances in neural information processing systems.* 2017, Bd. 30.

Lahoti, Preethi, et al. 2020. Fairness without demographics through adversarially reweighted learning. *Advances in neural information processing systems.* 2020.

Lee, Joshua K, et al. 2021. Fair selective classification via sufficiency. *International conference on machine learning.* 2021.

Li, Peizhao und Liu, Hongfu. 2022. Achieving fairness at no utility cost via data reweighing with influence. *International Conference on Machine Learning.* 2022.

Li, Yanhui, et al. 2022. Training data debugging for the fairness of machine learning software. *Proceedings of the 44th International Conference on Software Engineering.* 2022.

Liu, Suyun und Vicente, Luis Nunes. 2022. Accuracy and fairness trade-offs in machine learning: A stochastic multi-objective approach. *Computational Management Science.* 2022, Bd. 3, 19.

Lohia, Pranay K, et al. 2019. Bias mitigation post-processing for individual and group fairness. *IEEE International Conference on Acoustics, Speech and Signal Processing.* 2019.

Lu, Yuxiang , Cao, Shengcao und Wang, Yu-Xiong. 2025. Swiss Army Knife: Synergizing Biases in Knowledge from Vision Foundation Models for Multi-Task Learning. *The Thirteenth International Conference on Learning Representations.* 2025.

Luong, Binh Thanh, Ruggieri, Salvatore und Turini, Franco. 2011. k-NN as an implementation of situation testing for discrimination discovery and prevention. *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining.* 2011.

Madhavan, Poornima und Wiegmann, Douglas A. 2007. Similarities and differences between human--human and human--automation trust: an integrative review. *Theoretical Issues in Ergonomics Science.* 2007, Bd. 8.

Madras, David, Pitassi, Toni und Zemel, Richard. 2018. Predict responsibly: improving fairness and accuracy by learning to defer. *Advances in neural information processing systems.* 2018.

Marcinkevics, Ricards, Ozkan, Ece und Vogt, Julia E. 2022. Debiasing deep chest x-ray classifiers using intra-and post-processing methods. *Machine Learning for Healthcare Conference.* 2022.

Martinez, Natalia, Bertran, Martin und Sapiro, Guillermo. 2020. Minimax pareto fairness: A multi objective perspective. *International conference on machine learning.* 2020.

Mendler-Dünner, Celestine, Carovano, Gabriele und Hardt, Moritz. 2024. An engine not a camera: Measuring performative power of online search. *arXiv preprint arXiv:2405.19073.* 2024.

Menon, Aditya Krishna und Williamson, Robert C. 2018. The cost of fairness in binary classification. *Conference on Fairness, accountability and transparency.* 2018.

Noriega-Campero, Alejandro, et al. 2019. Active fairness in algorithmic decision making. *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society.* 2019.

Oneto, Luca, et al. 2019. Taking advantage of multitask learning for fair classification. *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society.* 2019.

Pedreschi, Dino, Ruggieri, Salvatore und Turini, Franco. 2009. Measuring discrimination in socially-sensitive decision records. *Proceedings of the 2009 SIAM international conference on data mining.* 2009.

Pleiss, Geoff, et al. 2017. On fairness and calibration. *Advances in neural information processing systems.* 2017.

Radcliffe, Thomas, Lockhart, Emily und Wetherington, James. 2024. Automated prompt engineering for semantic vulnerabilities in large language models. *Authorea Preprints.* 2024.

Roh, Yuji, et al. 2021. Sample selection for fair and robust training. *Advances in Neural Information Processing Systems.* 2021.

Savani, Yash, White, Colin und Govindarajulu, Naveen Sundar. 2020. Intra-processing methods for debiasing neural networks. *Advances in neural information processing systems.* 2020.

Stanczak, Karolina und Augenstein, Isabelle. 2021. A survey on gender bias in natural language processing. *arXiv preprint arXiv:2112.14168.* 2021.

Statistisches Bundesamt. Großstadtregionen im Wandel. [Online] Statistisches Bundesamt. [Zitat vom: 11. 02 2025.] https://www.destatis.de/DE/Themen/Querschnitt/Demografischer-Wandel/Aspekte/demografie-grossstadtregionen.html.

Strotherm, Janine, et al. 2024. Fairness in KI-Systemen. *Vertrauen in Künstliche Intelligenz: Eine multi-perspektivische Betrachtung.* 2024.

Suresh, Harini und Guttag, John V. 2019. A framework for understanding unintended consequences of machine learning. *arXiv preprint arXiv:1901.10002.* 2019.

Ustun, Berk, Liu, Yang und Parkes, David. 2019. Fairness without harm: Decoupled classifiers with preference guarantees. *International Conference on Machine Learning.* 2019.

Vincent, Nicholas, et al. 2021. Data leverage: A framework for empowering the public in its relationship with technology companies. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency.* 2021.

Wang, Hao, Ustun, Berk und Calmon, Flavio. 2019. Repairing without retraining: Avoiding disparate impact with counterfactual distributions. *International Conference on Machine Learning.* 2019.

Wang, Jialu, Wang, Xin Eric und Liu, Yang. 2022. Understanding instance-level impact of fairness constraints. *International Conference on Machine Learning.* 2022.

Wei, Dennis, Ramamurthy, Karthikeyan Natesan und Calmon, Flavio P. 2020. Optimized score transformation for fair classification. *Proceedings of Machine Learning Research.* 108, 2020.

Wikipedia. Simpson-Paradoxon. [Online] Wikipedia. [Zitat vom: 12. 02 2024.] https://de.wikipedia.org/wiki/Simpson-Paradoxon.

Woodworth, Blake, et al. 2017. Learning non-discriminatory predictors. *Conference on Learning Theory.* 2017.

Yerlikaya, Fahri Anil und Bahtiyar, Serif. 2022. Data poisoning attacks against machine learning algorithms. *Expert Systems with Applications.* 2022.

Zafar, Muhammad Bilal, et al. 2017. Fairness constraints: Mechanisms for fair classification. *International Conference on Artificial intelligence and statistics.* 2017.

Zemel, Rich, et al. 2013. Learning fair representations. *International conference on machine learning.* 2013.

Zhang, Bo, et al. 2020. Privacy for all: Demystify vulnerability disparity of differential privacy against membership inference attack. *arXiv preprint arXiv:2001.08855.* 2020.

Zhang, Brian Hu, Lemoine, Blake und Mitchell, Margaret. 2018. Mitigating unwanted biases with adversarial learning. *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society.* 2018.

Zliobaite, Indre, Kamiran, Faisal und Calders, Toon. 2011. Handling conditional discrimination. *IEEE 11th international conference on data mining.* 2011.