HMD (2022) 59:556–571
https://doi.org/10.1365/s40702-022-00844-1

SCHWERPUNKT

Der Einfluss von menschlichen Denkmustern auf künstliche Intelligenz – Eine strukturierte Untersuchung von kognitiven Verzerrungen

Lukas-Valentin Herm · Christian Janiesch · Patrick Fuchs

Eingegangen: 30. September 2021 / Angenommen: 1. Februar 2022 / Online publiziert: 2. März 2022

Zusammenfassung Künstliche Intelligenz (KI) dringt vermehrt in sensible Bereiche des alltäglichen menschlichen Lebens ein. Es werden nicht mehr nur noch einfache Entscheidungen durch intelligente Systeme getroffen, sondern zunehmend auch komplexe Entscheidungen. So entscheiden z. B. intelligente Systeme, ob Bewerber in ein Unternehmen eingestellt werden sollen oder nicht. Oftmals kann die zugrundeliegende Entscheidungsfindung nur schwer nachvollzogen werden und ungerechtfertigte Entscheidungen können dadurch unerkannt bleiben, weshalb die Implementierung einer solchen KI auch häufig als sogenannte Blackbox bezeichnet wird. Folglich steigt die Bedrohung, durch unfaire und diskriminierende Entscheidungen einer KI benachteiligt behandelt zu werden. Resultieren diese Verzerrungen aus menschlichen Handlungen und Denkmustern spricht man von einer kognitiven Verzerrung oder einem kognitiven Bias. Aufgrund der Neuigkeit dieser Thematik ist jedoch bisher nicht ersichtlich, welche verschiedenen kognitiven Bias innerhalb eines KI-Projektes auftreten können. Ziel dieses Beitrages ist es, anhand einer strukturierten Literaturanalyse, eine gesamtheitliche Darstellung zu ermöglichen. Die gewonnenen Erkenntnisse werden anhand des in der Praxis weit verbreiten Cross-Industry Standard Process for Data Mining (CRISP-DM) Modell aufgearbeitet und klassifiziert. Diese Betrachtung zeigt, dass der menschliche Einfluss auf eine KI in

Lukas-Valentin Herm (✉)
Julius-Maximilians-Universität Würzburg, Sanderring 2, 97070 Würzburg, Deutschland
E-Mail: lukas-valentin.herm@uni-wuerzburg.de

Christian Janiesch
TU Dortmund University, Dortmund, Deutschland
E-Mail: christian.janiesch@tu-dortmund.de

Patrick Fuchs
Julius-Maximilians-Universität Würzburg, Sanderring 2, 97070 Würzburg, Deutschland
E-Mail: patrick.fuchs@uni-wuerzburg.de

jeder Entwicklungsphase des Modells gegeben ist und es daher wichtig ist „menschähnlichen“ Bias in einer KI explizit zu untersuchen.

Schlüsselwörter Menschliche Denkmuster · Maschinelles Lernen · Künstliche Intelligenz · Literaturanalyse

The Impact of Human Thinking on Artificial Intelligence—A Structured Investigation of Cognitive Biases

Abstract Artificial intelligence (AI) is increasingly penetrating sensitive areas of everyday human life, resulting in the ability to support humans in complex and difficult tasks. The result is that intelligent systems are capable of handling not only simple but also complex tasks. For example, this includes deciding whether an applicant should be hired or not. Oftentimes, this decision-making can be difficult to comprehend, and consequently incorrect decisions may remain undetected, which is why these implementations are often referred to as a so-called black box. Consequently, there is the threat of unfair and discriminatory decisions by an intelligent system. If these distortions result from human actions and thought patterns, it is referred to as a cognitive bias. However, due to the novelty of this subject, it is not yet apparent which different cognitive biases can occur within an AI project. The aim of this paper is to provide a holistic view through a structured literature review. Our insights are processed and classified according to the Cross-Industry Standard Process for Data Mining (CRISP-DM) model, which is widely used in practice. This review reveals that human influence on an AI is present in every stage of the model’s development process and that “human-like” biases in an AI must be examined explicitly.

Keywords Cognitive Biases · Machine Learning · Artificial Intelligence · Literature Review

1 Einleitung

Künstliche Intelligenz (KI) prägt zunehmend das Leben unserer Zeit. Durch die vielseitigen Einsatzgebiete der Technologie sowie ihre stetige Weiterentwicklung übernehmen intelligente Systeme vermehrt ursprünglich menschliche Entscheidungen und beeinflussen somit das tägliche Leben vieler Menschen sowohl bewusst als auch unbewusst (Zhang et al. 2018). Dabei erstellen diese Algorithmen Empfehlungen, um Menschen in verschiedensten Anwendungsfällen, direkt oder indirekt, zu unterstützen. Jedoch basieren Entscheidungen von KI in vielen Fällen auf identifizierten Mustern und Zusammenhängen in großen Datenmengen (Murphy 2012). Dies resultiert in einer zunehmenden Intransparenz und damit einer fehlenden Nachvollziehbarkeit, sodass selbst Entwickler die innere Entscheidungslogik dieser intelligenten Systeme nur noch schwer verstehen können. In diesem Zusammenhang wird KI häufig auch als *Blackbox* bezeichnet (Herm et al. 2021).

Sind die Entscheidungen eines Systems für den Menschen nicht nachvollziehbar, bleibt ungewiss, wieso ein System diese Entscheidungen getroffen hat. Diese Problematik hat insbesondere zur Folge, dass Endanwender und Unternehmen in der Adaption dieser intelligenten Systeme zögerlich agieren, da sie nicht sicherstellen können, ob von Dritten etwaige Fehler und Verzerrungen im Entscheidungsprozess entstanden sind. Entscheidet eine KI beispielsweise nicht anhand von objektiven Fakten, sondern diskriminierend aufgrund von verzerrten Fakten, kann dies unbemerkt bleiben. Verschiedenste aktuelle Beispiele, wie die Benachteiligung von Personengruppen (Udeshi et al. 2018), verdeutlichen, dass eine KI vergleichbar mit menschlicher Intelligenz entscheidet und es zu befangenen Entscheidungen kommen kann, wenn die Datengrundlage nicht sachliche Verzerrungen aufweist. Die Entscheidungsfindung einer KI kann somit Anzeichen menschlicher Denkmuster aufweisen. Die Entscheidung ist daher nach objektiven Gesichtspunkten nicht fair und weist eine inhärente Neigung des maschinellen Modells auf, die dazu führt, dass eine KI systematisch falsche Ergebnisse erzielt (Cramer et al. 2018). Eine KI und deren Entscheidungsprozess enthält somit einen *Bias* (Sengupta et al. 2018).

Die Identifikation und Vermeidung von Bias in einer KI sind Themen, die in den vergangenen Jahren vermehrt in den Fokus der Forschung gerückt sind. Dabei handelt sich vergleichbar mit der Untersuchung menschlichen Denkmustern um ein vielschichtiges Forschungsgebiet (Hundman et al. 2018). Gleichwohl ist der Entwicklungsprozess einer KI mehrstufig und komplex, sodass in jeder Stufe der Entwicklung ein Bias entstehen kann. Bias in einer KI können insofern nicht nur auf einem zuvor existenten Bias im verwendeten Trainingsdatensatz zurückgeführt werden, sondern auch auf die menschliche Beteiligung in jeder Phase der Entwicklung. Eine Übertragung bestehender sozialer Verzerrungen und menschlicher Denkmuster auf eine KI ist infolgedessen grundsätzlich in jedem Entwicklungsschritt möglich (Garcia-Gathright et al. 2018). In diesem Zuge versuchen Forschungsdomänen wie *erklärbare künstliche Intelligenz* (engl. explainable artificial intelligence, XAI) die Probleme in der Nachvollziehbarkeit von Entscheidungen von KI zu adressieren und die Entscheidungsprozesse für den Menschen transparent und nachvollziehbar zu gestalten (Herm et al. 2021).

Während diese Thematik vermehrt in den Fokus der Forschung und Praxis rückt, fehlt es jedoch aktuell an einer gesamtheitlichen Darstellung potenzieller menschlicher Befangenheiten bei der Anwendung von KI (Hellström et al. 2020, Mehrabi et al. 2021). Daher ist es Ziel dieses Beitrages zu untersuchen, wie sich menschliche Denkmuster im Entwicklungsprozess auf eine KI übertragen und deren Entscheidungen beeinflussen können. Daraus leitet sich folgende Forschungsfrage ab:

FF: *Wie spiegeln sich menschliche Denkmuster in den Entscheidungen einer künstlichen Intelligenz wider?*

Grundlage für die Beantwortung der Forschungsfrage bildet eine umfangreiche Literaturanalyse nach dem Leitfaden von vom Brocke et al. (2015). Aufbauend auf den Erkenntnissen der Literaturanalyse wird eine Betrachtung möglicher Ursachen für Bias in einer KI anhand des weit verbreiteten Vorgehensmodells *Cross-Industry Standard Process for Data Mining* (CRISP-DM) vorgenommen (Chapman et al. 2000). Der vorliegende Beitrag strukturiert sich dabei wie folgt: Im 2. Kapitel werden

die notwendigen theoretischen Grundlagen erläutert. Anschließend, findet in Kap. 3 eine Übersicht sowie Erläuterung der verschiedenen Bias statt, welche in Kap. 4 kritisch gewürdigt werden. Eine Schlussfolgerung, Limitationen und Ausblick in Kap. 5 runden den Beitrag ab.

2 Theoretische Grundlagen

2.1 Kognitive Informationsverarbeitung und resultierende Bias

Unter dem Begriff der Kognition werden in der kognitiven Psychologie alle informationsverarbeitenden Prozesse eines intelligenten Systems verstanden. Hierzu zählen beispielsweise mentale Prozesse, wie das Denken, das Sprachverständnis und das Problemlösen. Sie folgen den elementaren Phasen der Informationsverarbeitung (Hundman et al. 2018). Diese beginnt mit der Verarbeitung (Encodierung) der Informationen und verläuft über deren Aufbewahrung (Speicherung als mentale Repräsentation) bis zur deren Wiedergabe (Aufruf) zu einem späteren Zeitpunkt (Raab et al. 2010). Jedoch ist die Kapazität der menschlichen Informationsverarbeitung begrenzt. Dies kann zu einer selektiven Aufnahme und verzerrten Verarbeitung der verfügbaren Informationen führen. Dies resultiert aus der Absicht, schnelle und adäquate Reaktionen anhand einer Entscheidung zu ermöglichen, indem Informationen mit möglichst geringem Aufwand verarbeitet werden (Cramer et al. 2018). Tritt folglich bei der menschlichen Informationsverarbeitung eine verzerrte, selektive Wahrnehmung auf, die zu systematischen Urteilsfehlern in der Entscheidungsfindung führt, ist in der Psychologie die Sprache von kognitiven Bias. Bei solchen Urteilsfehlern handelt es sich nicht um zufällige Diskrepanzen im Entscheidungsprozess, sondern um eine systematische Befangenheit innerhalb der Informationsverarbeitung. Diese Denkschemata sind Bestandteil der menschlichen Informationsverarbeitung und können zu unbewussten Verschiebungen in der Gewichtung von Entscheidungsfaktoren führen (Raab et al. 2010).

2.2 Bias in Künstlicher Intelligenz

Im Forschungsbereich Wirtschaftsinformatik wird der Term KI als allgemeiner Ausdruck für intelligente Agenten verstanden, welche auf Basis von datenbasierten Beobachtungen Entscheidungslogik und Wissen generieren (Herm et al. 2021). Dabei benötigen die intelligenten Agenten kognitive Fähigkeiten, um Muster und Problemstellungen, ähnlich wie bei Menschen, selbstständig zu erkennen. In den letzten Jahren hat insbesondere die Subklasse *maschinelles Lernen* (ML) viel Aufmerksamkeit erlangt. Hierbei werden anhand von mathematischen Algorithmen Modelle trainiert, die anhand von empirischen Daten nicht lineare Beziehungen erkennen und daraus Entscheidungsmuster entwickeln (Janiesch et al. 2021). Das *tiefe Lernen* (engl. deep learning, DL) ist, wiederum ein Teilgebiet des maschinellen Lernens, bei welchem primär tiefe künstliche neuronale Netze eingesetzt werden. Der Aufbau eines solchen Netzes ähnelt in seiner Grundstruktur und Funktionsweise einem menschlichen Gehirn und ist damit fähig, hochpräzise Vorhersagen zu treffen (Zhang et al. 2018).

Infolge des Einflusses, den eine KI auf einzelne Menschen haben kann, ist ein verantwortungsvoller Umgang und Einsatz der KI unerlässlich. Dies betrifft sowohl den Umgang des Menschen mit den Systemen als auch die Entwicklung dieser Systeme (Henderson et al. 2018). Um dies zu wahren, ist es notwendig, dass die genutzten KI-Implementierungen die Privatsphäre und die menschlichen Werte der Nutzer, in einer kontrollierbaren und einsehbaren Art und Weise respektieren (Fjeld et al. 2020). Sollten diese Herausforderungen nicht bei der Entwicklung solcher intelligenten Systeme adressiert werden können, kann es zu Bias kommen (Udeshi et al. 2018). Schlussendlich bewirken die Bias eine Ungleichverteilung innerhalb des genutzten ML-Modells. Durch die grundlegende Eigenschaft, auf Trainingsdaten zu lernen, können diese Modelle daher bestehende Verzerrungen in der Gesellschaft bewusst oder unbewusst aufnehmen, replizieren und sogar verstärken (Garcia-Gathright et al. 2018). So, identifizieren ML-Modelle in der Trainingsphase Muster und Zusammenhänge, welche anschließend in der Phase der Entscheidungsfindung in die Ergebnisse des Modells einfließen (Cramer et al. 2018).

Bias innerhalb intelligenter Systeme zeigt sich häufig in Form von Benachteiligung bestimmter Personengruppen auf Basis ihrer sensiblen Attribute, wie zum Beispiel ihrer Religion, ihrer Rasse, ihres Alters oder ihres Geschlechts (Udeshi et al. 2018). So enthielt ein von Amazon.com, Inc. entwickeltes intelligentes System, welches eine Vorauswahl von potenziellen Bewerbern durchführen sollte, einen Geschlechter-Bias. Dies resultierte aus historischen Daten, da Bewerber überwiegend männlich waren. Somit erfolgte somit eine Reproduktion und Verstärkung von Ungleichheiten (Nier 2018). Da jedoch solche Komplikationen grundsätzlich in jedem Entwicklungs- und Anwendungsschritt entstehen können, ist nicht eindeutig, welche verschiedenen Formen von Bias insgesamt existieren.

3 Übersicht der kognitiven Bias im Kontext des CRISP-DM

3.1 Vorgehen

Vorgehen der strukturierten Literaturanalyse Die Grundlage für die Analyse bildet eine strukturierte Literaturanalyse nach vom Brocke et al. (2015), durch welche bestehenden Probleme aus der Forschung und Anwendung zu Bias in KI und damit der Implementierung von ML-Modellen aufgezeigt werden soll. Um eine repräsentative Aufarbeitung der Thematik zu ermöglichen, werden sowohl wissenschaftliche Datenbanken aus der Wirtschaftsinformatik (AIS eLibrary, ScienceDirect) und der Informatik (IEEE Xplore, ACM Digital Library) untersucht. Hierfür wurde der folgende Suchterm benutzt *„machine learning AND bias“*, welcher 1644 potenzielle Forschungsbeiträge lieferte. In Anlehnung an vom Brocke et al. (2015), wurde auf Grund der Aktualität der Thematik keine Eingrenzung anhand von Rankings vorgenommen. Anhand einer Titel- und Abstract-Analyse, einer Volltextanalyse sowie einer Vorwärts- und Rückwärtssuche wurden schlussendlich 259 Forschungsbeiträge berücksichtigt. Für die Analyse dieser Beiträge wurden alle Bias erhoben und thematisch ähnliche Bias zusammengefasst.

Einordnung in CRISP-DM Referenzmodell Das CRISP-DM, ist ein standardisiertes und etabliertes Vorgehensmodell für Data-Mining-Projekte aus dem Jahre 1996, welches sich innerhalb KI- bzw. ML-Projekten als Vorgehensmodell durchgesetzt hat (Wiemer et al. 2019). Neben einem definierten Benutzerhandbuch findet sich innerhalb des Standards ein Vorgehensmodell, welches die verschiedenen Phasen innerhalb der Implementierung und Nutzung von ML-Modellen beschreibt. Dieses Referenzmodell umfasst sechs Phasen (Chapman et al. 2000):

1. Geschäftsverständnis: In der ersten Phase, wird die Problemstellung aus betriebswirtschaftlicher Perspektive betrachtet und analysiert, um daraus Ziele zu entwickeln.
2. Datenverständnis: Hier wird die Datengrundlage erhoben, die Datenqualität erfasst sowie die Daten untersucht.
3. Datenaufbereitung: Dient zur Auswahl, Bereinigung, Transformation und Formatierung der Daten in ein passendes Format.
4. Modellierung: Evaluierung und Parametrisierung verschiedener Modelle durch einen iterativen Prozess.
5. Evaluierung: Evaluierung der Modelle im produktiven Einsatz.
6. Bereitstellung: Erkenntnisse werden geordnet, präsentiert und genutzt.

Die verschiedenen Phasen sowie die darin enthaltenen Aufgaben sind in der Abb. 1 dargestellt. Zusätzlich werden zur Übersicht die potenziellen Bias, welche in den verschiedenen Phasen auftreten können, genannt. Eine nach Aufgaben gegliederte Auflistung findet sich in Tab. 1 am Ende des Kapitels.

3.2 Geschäftsverständnis

Die Phase *Geschäftsverständnis* umfasst die Aufgaben Projektdefinition, Situationsbewertung, Analyseziele und Projektplan. Die dazugehörigen Bias werden im Folgenden beschrieben.

Projektdefinition Innerhalb der ersten Phase werden die Prioritäten des Auftraggebers definiert und damit Ziele angestrebt. Die zu erreichenden Ziele können in einem *Funding Bias* resultieren, da mögliche Ziele nicht objektiv erarbeitet, sondern durch die Stakeholder des Projektes beeinflusst werden (Mehrabi et al. 2021). Eine mögliche Ursache für einen Funding Bias in der Projektdefinition eines ML-Projektes ist die Tendenz zur bevorzugten Aufnahme von Informationen, die die eigenen Einstellungen bestärken.

Situationsbewertung In der Situationsbewertung werden die gesammelten Informationen aus dem vorangegangenen Schritt aufgearbeitet und anhand der verfügbaren Ressourcen, bestehenden Restriktionen erfasst sowie ein Projektteam zusammengestellt (Chapman et al. 2000). Die unterschiedlichen Erfahrungslevel und kognitiven Bias der Projektbeteiligten spiegeln sich durch deren Beteiligung in jedem Entwicklungsschritt wider. Des Weiteren können durch Wissens- und Erfahrungslücken innerhalb des Projektteams wichtige Aspekte unberücksichtigt bleiben und

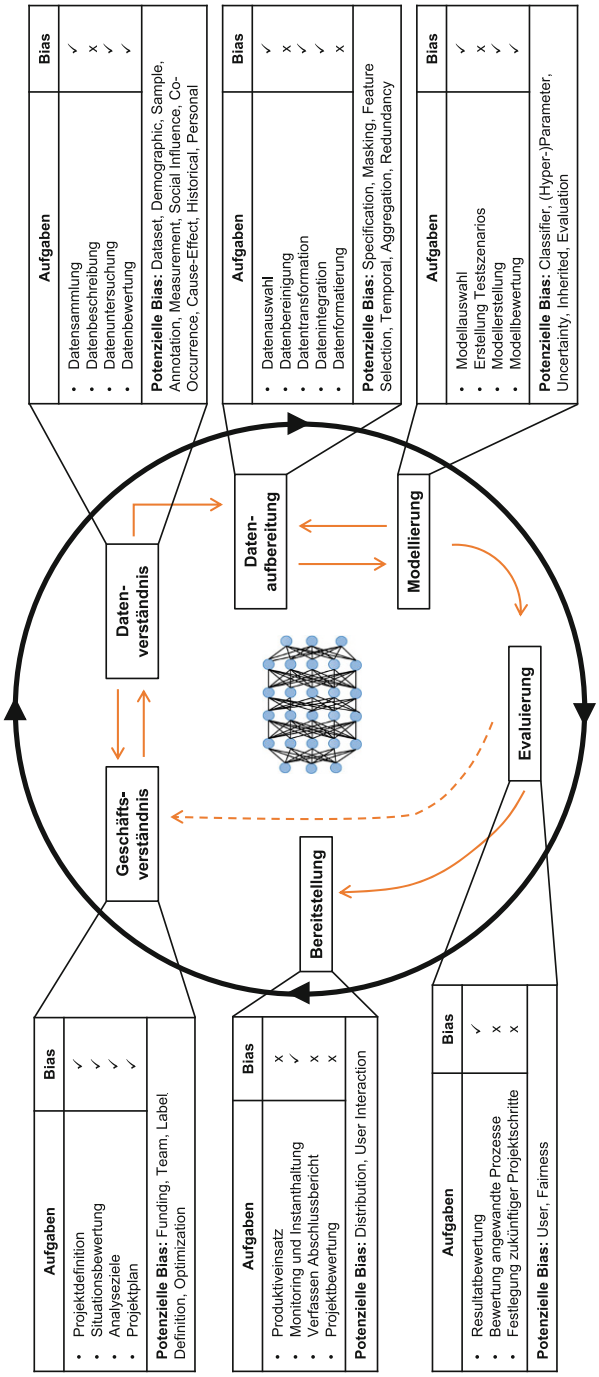

Abb. 1 Phasen und Aufgaben des CRISP-DM sowie potenzielle Bias innerhalb von KI-Projekten

bestimmte Probleme nicht erkannt werden. Diese Problematik wird auch als *Team Bias* verstanden (Cramer et al. 2018).

Analyseziele Bei der Bestimmung der Analyseziele werden die betriebswirtschaftlichen Geschäftsziele, in eine technische Beschreibung überführt (Chapman et al. 2000). Bei dieser Überführung kann es durch eine ungenaue Beschreibung der angestrebten Ergebnisse zu inkorrekten Vorhersagen durch das Modell kommen. Diese Bias werden als *Label Definition Bias* bezeichnet (Mujtaba und Mahapatra 2019). Bei der Übersetzung der betriebswirtschaftlichen Zielsetzung in technische Analyseziele muss eine Priorisierung der Ziele erfolgen. Für einzelne Vorhaben müssen Kompromisse gefunden werden. Ein Mittelweg aus einer allgemein hohen Vorhersagequalität und einer fairen Behandlung einzelner Gruppen ist notwendig. Hierfür existieren verschiedene statische Metriken für die Definition von Fairness, weshalb durch diese inhärente Mehrdeutigkeit, die Bestimmung des Mittelweges ein sogenannter *Optimization Bias* entstehen kann (Dobbe et al. 2018).

Projektplan Bei der Erstellung des Projektplanes werden die zuvor erreichten Schritte verschriftlicht, weshalb die eben beschriebenen Funding, Label Definition und Optimization Bias unbemerkt übernommen werden können (Chapman et al. 2000, Cramer et al. 2018).

3.3 Datenverständnis

In dieser Phase erfolgten die Datensammlung, Datenbeschreibung, Datenuntersuchung und Datenbewertung. Da für die Datenbeschreibung keine Bias identifiziert werden konnten, wird diese Aufgabe im Folgenden nicht weiter betrachtet.

Datensammlung Die Phase des Datenverständnisses beginnt mit der Beschaffung der benötigten Daten für die Entwicklung eines ML-Modells (Chapman et al. 2000). Wird für das Training des ML-Modells ein Datensatz verwendet, der die Umwelt nicht umfangreich und korrekt abbildet, kann man das Modell nicht angemessen verallgemeinern. Diese Problematik wird als *Dataset Bias* bezeichnet. So ein nichtrepräsentativer Datensatz kann durch menschliche und systemische Faktoren entstehen und zu einer Diskrepanz in den Verteilungen hinsichtlich der Demographie führen (*Demographic Bias*) (Li und Vasconcelos 2019). Wird beispielsweise ein ML-Modell für die Identifikation von Gesichtern entwickelt, wird ein Trainingsdatensatz benötigt, welcher alle existierenden Gesichtsstrukturen und -formen enthält. Durch die nicht zufällige und von ethnischen sowie geografischen Gegebenheiten abhängige Zusammenstellung des Trainingsdatensatzes wird ein ML-Modell in den Trainingsdaten nicht enthaltene Gesichtsstrukturen später allerdings unter Umständen nicht korrekt identifizieren können (Sengupta et al. 2018). Ein sogenannter *Sample Bias* entsteht somit, durch eine nicht zufällige Auswahl von Untergruppen, welches sich durch eine Unter- oder Überrepräsentation einzelner Beobachtungen aus einem Segment äußert (Hellström et al. 2020). Diese kann neben der ungleichgewichtigen Berücksichtigung von Merkmalswerten oder Klassen auch Populationen betreffen, welche absichtlich oder unabsichtlich durch zu geringe Stichproben oder

Selektionsverzerrungen entstehen können (Li und Vasconcelos 2019). Des Weiteren werden Daten oftmals von Crowdworkern oder Projektmitgliedern gesammelt und klassifiziert (Hube et al. 2019). In einem solchen Vorgang können Menschen ihre Vorurteile auf die Daten und infolgedessen auf das trainierte ML-Modell übertragen (Hellström et al. 2020), welches auch als *Annotation Bias* bezeichnet wird (Hube et al. 2019). Dabei kann es zu systematischen Messfehlern innerhalb der Beobachtungen kommen (*Measurement Bias*) (Hellström et al. 2020). Dies betrifft auch die Erhebung von Bewertungsdaten, da Personen sich oftmals durch andere Bewertungen oder Gruppen bei ihrer eigenen Bewertung beeinflussen lassen (*Social Influence Bias*) (Mehrabi et al. 2021). Schlussendlich kann es auch zu einem Verhaltenswechsel innerhalb der erhobenen Daten kommen, weshalb sich Daten, die über eine längere Zeitperiode erhoben werden, sich substanziell unterscheiden. Hier spricht man von einem *Temporal Bias* (Cramer et al. 2018).

Datenuntersuchung Während dieser Aufgabe werden Abfrage-, Visualisierungs- und Reporting-Verfahren eingesetzt, um ersten Fragen des Data-Mining beziehungsweise des ML-Projektes zu beantworten (Chapman et al. 2000). Treten in den Trainingsdaten einzelne Werte überproportional häufig in Kombination auf, kann ein *Co-Occurance Bias* zu falschen Rückschlüssen einer KI führen. Dies lässt sich auf die Lernverfahren von ML-Modellen und deren Wissensaufbau basierend auf existierenden Korrelationen in den Trainingsdaten zurückführen (Hellström et al. 2020). Anschließend kann es bei der Analyse der Daten zu einem *Cause-Effect Bias* und damit zu einem Trugschluss kommen. Nach diesem Trugschluss wird Kausalität durch Korrelation impliziert und damit fälschliche Annahmen getroffen (Mehrabi et al. 2021).

Datenbewertung In der letzten Aufgabe werden die Daten hinsichtlich ihrer Qualität genauer untersucht (Chapman et al. 2000). Wird hier die erhobene Umwelt in ihrer tatsächlichen Form in den Daten abgebildet, kann dies zu einem *Historical Bias* im darauf trainierten ML-Modell führen (Suresh und Guttag 2019). Existierende Verzerrungen und sozio-technische Probleme der Umwelt werden vom Modell imitiert. Die bestehenden Diskrepanzen zwischen der Umwelt wie sie ist und den moralisch und sozial angestrebten Werten, werden vom Modell kodiert und finden sich in den Entscheidungen des Modells wieder. Ebenso kann es zu persönlichen Befangenheiten, auch *Personal Bias* genannt, innerhalb der Evaluierung kommen, welches auch die Benach- oder Bevorteilung von demographischen Gruppen bewirken kann (Sharma und Wehrheim 2019).

3.4 Datenaufbereitung

Es finden die Datenauswahl, Datenbereinigung, Datentransformation, Datenintegration und Datenformatierung statt. Wir konnten für die Aufgaben Datenbereinigung und Datenformatierung keine Bias identifizieren.

Datenauswahl Während der ersten Aufgabe dieser Phase entscheiden die Projektbeteiligten, welche Daten verwendet werden sollen (Chapman et al. 2000). Dabei

erfolgt die Auswahl durch die Projektbeteiligten und erfordert ein gutes Verständnis der Problemstellung sowie der Zielsetzung des Projektes, da es ansonsten zu Verzerrungen in der Auswahl kommen kann (*Specification Bias*) (Hellström et al. 2020). Sollten sensible Daten anonymisiert werden müssen, kann eine fehlerhafte Kodierung zu einem *Masking Bias* führen (Mujtaba und Mahapatra 2019). Ist der Auswahlprozess der Features für das ML-Modell die Ursache für Verzerrungen, findet der Begriff *Feature Selection Bias* Anwendung (Jensen und Neville 2002). Bei der Auswahl der Daten kann es dazu kommen, dass die Beziehung zwischen einzelnen Datenobjekten und der Zielausgabe, d. h. den Labels, in den ausgewählten Daten stärker ausgeprägt abgebildet wird, als es der tatsächliche Zusammenhang in der Realität ist. Dementsprechend kann sich die Auswahl der Features negativ auf die Vorhersagegenauigkeit und die Verallgemeinerungsfähigkeit des späteren Modells auswirken. Ähnlich wie bei dem Sammeln der Daten in der Phase des Datenverständnis, kann die Berücksichtigung von Daten aus verschiedenen Zeitpunkten zu einem *Temporal Bias* führen (Cramer et al. 2018).

Datentransformation Bestandteil der Aufgabe ist die Durchführung konstruktiver Datenaufbereitungs-operationen (Chapman et al. 2000). Bei dieser Aufgabe kann es bei der Ableitung neuer Werte durch eine inadäquate Zusammenfassung bestimmter Untergruppen zu einem *Aggregation Bias* kommen. Dieser Bias tritt auf, wenn unterschiedliche Populationen in unangemessener Weise kombiniert werden (Suresh und Guttag 2019).

Datenintegration Bei der Datenintegration werden mehrere Datensätze miteinander kombiniert, um einen umfassenden Datensatz für die Entwicklung eines ML-Modells zu erstellen (Chapman et al. 2000). Bei der Zusammenführung verschiedener Datensätze muss darauf geachtet werden, dass Verzerrungen durch ein Mehrfachauftreten von Objekten in den einzelnen Datensätzen entstehen können. Diese Verzerrungen werden unter dem Begriff des *Redundancy Bias* zusammengefasst (Cramer et al. 2018).

3.5 Modellierung

Es wird die Modellauswahl, die Erstellung des Testszenarios, die Modellerstellung und die Modellbewertung vorgenommen. Für die Erstellung des Testszenarios wurden keinerlei Bias festgestellt.

Modellauswahl Die Phase der Modellierung beginnt mit der Auswahl der Modellierungstechnik und damit auf welchem Algorithmus ein ML-Modell aufgebaut werden soll (Chapman et al. 2000). Hierbei zeigen bestehende Forschungsarbeiten, dass die Wahl des Algorithmus signifikante Auswirkung auf Wirksamkeit des ML-Modells haben kann. Solche Verzerrungen werden als *Classifier Bias* bezeichnet und resultieren aus den spezifischen Eigenschaften (z. B. der Lernstrategie) der verschiedenen ML-Modelle (Tang und Liu 2005).

Modellerstellung Bei der Erstellung des Modells werden verschiedene Varianten des ML-Modells auf dem Trainingsdatensatz trainiert und Justierungen bei den zugehörigen Parametern vorgenommen (Chapman et al. 2000). Insbesondere bei Algorithmen des DL müssen eine Vielzahl an Modellparametern, die sogenannten Hyperparameter, an die Problemstellung angepasst werden, da sonst das genutzte Modell nicht auf den zu trainierenden Sachverhalt passt und damit nicht präzise trainieren kann (Zhang et al. 2018). Diese lernt das Modell nicht automatisiert während des Trainings auf dem Trainingsdatensatz, sondern diese müssen durch die Projektbeteiligten aktiv (nach einer Parametersuche) händisch ausgewählt und eingestellt werden. Entstehen hier Verzerrungen, spricht man von einem *(Hyper-)Parameter Bias* (Hellström et al. 2020). Nahe damit verbunden ist die definierte Höhe des Unsicherheitswertes, bei dem ein ML-Modell eine Empfehlung ausspricht. Dies wird durch die Projektbeteiligten definiert und kann so durch subjektive Wahrnehmung zu einem *Uncertainty Bias* führen. Sollten bestehende ML-Modelle wiederverwendet werden, können bestehende Bias ebenfalls in einen Kontext übertragen werden (*Inherited Bias*) (Hellström et al. 2020).

Modellbewertung Die letzte Aufgabe der Modellierung zielt auf die Bewertung des entwickelten Modells ab. Die Projektmitglieder greifen auf ihre Erfahrung, die festgelegten Erfolgskriterien sowie das zuvor festgelegte Testszenario zurück (Chapman et al. 2000). Hierzu wird auf bestehende Kennzahlen und Bewertungsmetriken zurückgegriffen. Unterlaufen hier Fehler, spricht man von einem *Evaluation Bias* (Suresh und Guttag 2019).

3.6 Evaluierung

Während die Phase der Evaluierung aus mehreren Aufgaben besteht, wurden nur Bias mit der Aufgabe Resultatbewertung festgestellt. Die weitere Betrachtung der Bewertung angewandte Prozesse und Festlegung über zukünftige Projektschritte findet daher nicht statt.

Resultatbewertung Das Ziel in dieser Aufgabe ist es zu bewerten, inwieweit das ML-Modell die Geschäftsziele erfüllt (Chapman et al. 2000). Werden die Resultate des ML-Modell zusätzlich zur vorangegangenen Bewertung des Modells in einem produktiven Szenario betrachtet, kann es bei diesem Schritt durch die beteiligten Testpersonen zu einem *User Bias* kommen, da diese nicht uneingeschränkt objektiv sind und somit persönlichen Bias bewusst oder unbewusst in die Bewertung des Systems einfließen lassen (Lutfi et al. 2013). In sensiblen Einsatzgebieten ist vor allem die Fairness des ML-Modells zu berücksichtigen, um sozial oder moralisch bedenkliche Entscheidungen (*Fairness Bias*) zu vermeiden (Mehrabi et al. 2021).

3.7 Bereitstellung

Ähnlich wie bei der Evaluierungsphase wurde bei der Bereitstellung nur Bias innerhalb einer Aufgabe, Monitoring und Instanthaltung, festgestellt. Für Produktiveinsatz, Verfassen Abschlussbericht und Projektbewertung wurden keine Bias aufgeführt.

Monitoring und Instanthaltung Monitoring und Instandhaltung sind Faktoren für einen erfolgreichen Einsatz eines ML-Modells (Chapman et al. 2000). ML-Modelle werden häufig auf Daten, welche aus mehreren Quellen stammen, trainiert. Wird das Modell trotzdem in einer neuen und unbekannten Domäne genutzt, kann daraus ein *Distribution Bias* resultieren. Die Qualität der Ergebnisse des Modells ist folglich nicht gesichert (Erfani et al. 2016). Ebenso können sich Verzerrungen bei der Interaktion mit dem Nutzer ergeben. Dies tritt auf, wenn der Nutzer durch das Interface gelenkt wird und wird daher als *User Interaction Bias* bezeichnet (Heindorf et al. 2019).

3.8 Zusammenfassung der Bias

Basierend auf den Ergebnissen aus dem vorherigen Kapitel, fasst die folgende Tabelle die verschiedenen Bias phasenübergreifend zusammen.

Tab. 1 Übersicht CRISP-DM-Aufgaben und potenzielle Bias

Phase	Aufgabe	Bias
Geschäftsverständnis	Projektdefinition	Funding Bias
	Situationsbewertung	Team Bias
	Analyseziele	Label Definition Bias, Optimization Bias
	Projektplan	Funding, Label Definition, Optimization Bias
Datenverständnis	Datensammlung	Dataset Bias, Demographic Bias, Sample Bias, Annotation Bias, Measurement Bias, Social Influence Bias, Temporal Bias
	Datenuntersuchung	Co-Occurrence Bias, Cause-Effect Bias
	Datenbewertung	Historical Bias, Personal Bias
Datenaufbereitung	Datenauswahl	Specification Bias, Masking Bias, Feature Selection Bias, Temporal Bias
	Datentransformation	Aggregation Bias
	Datenintegration	Redundancy Bias
Modellierung	Modellauswahl	Classifier Bias
	Modellerstellung	(Hyper-)Parameter Bias, Uncertainty Bias, Inherited Bias
	Modellbewertung	Evaluation Bias
Evaluierung	Resultatbewertung	User Bias, Fairness Bias
Bereitstellung	Monitoring und Instanthaltung	Distribution Bias, User Interaction Bias

4 Implikationen für die Umsetzung von KI-Projekten

Die Ergebnisse der strukturierten Literaturanalyse zeigen eine diffuse und fragmentierte Menge an potenziellen Bias innerhalb von KI- bzw. ML-Projekten. Dementsprechend werden verschiedene Herangehensweisen zur Vermeidung und Behebung dieser Bias vorgeschlagen, weshalb sich keine einheitliche Vorgehensweise ableiten lassen kann. Dennoch ist zuerst die Identifikation der Bias sowie deren mögliche Ursachen unerlässlich (Holstein et al. 2019). Innerhalb unserer Forschung haben wir zwei übergeordnete Ansätze synthetisieren können, welche es Entwicklern und Anwendern anhand von Methoden und Werkzeuge ermöglichen diese Bias zu identifizieren:

1. Durch eine explorative Datenanalyse können Probleme wie sensible Attribute oder unterpräsentierte Gruppen erkannt und durch Transformationstechniken ausgeglichen werden, bevor diese für das Training von ML-Modellen genutzt werden. Gleichzeitig kann hier durch eine überarbeitete Datensammlung angestoßen werden.
2. Durch den Einsatz von XAI können ML-Modelle nachvollziehbar gemacht werden und somit den Einfluss von Bias auf die inhärenten Entscheidungslogik der ML-Modelle darstellen. Zum einen ermöglicht dies eine kritische Überprüfung und damit eine iterative Anpassung der ML-Modelle. Zum anderen kann im Nachgang eine Evaluierung der implementierten ML-Modelle vorgenommen werden, welche im Zuge des CRISP-DM Vorgehens eine erneute Anpassung der Datengrundlage oder der Modellentwicklung ermöglicht.

Gleichzeitig stellten wir innerhalb unserer Untersuchung fest, dass es an rechtlichen Rahmenwerken und Orientierungspunkten fehlt (Hanika 2019), da potenzielle Verzerrungen durch Bias zu jedem Zeitpunkt bestehen können. Dem einhergehend bestehen auch sozio-technische Unsicherheiten seitens der Entwickler und Anwender, welche eine uneingeschränkte KI-Adoption in der Praxis behindern (Wanner et al. 2021). Somit ist eine permanente und kritische Betrachtung dieser Systeme unabdingbar. Dennoch ist zu erwähnen, dass die Befolgung dieser Ansätze nicht automatisch zum Erfolg führen muss, da zum einen je nach Anwendungskontext sich weitere und damit neue Bias ergeben können und zum anderen Beiträge wie Slack et al. (2020) bereits Schwächen innerhalb geläufiger XAI-Werkzeuge aufgezeigt haben.

5 Schlussfolgerung, Limitationen und Ausblick

KI und vor allem ML-Modelle spielen in der Gesellschaft eine immer gewichtigere Rolle (Zhang et al. 2018). Über intelligente Systeme übernehmen sie nicht mehr nur noch einfache Entscheidungen, sondern auch verstärkt komplexere Aufgaben. Die Menge an potenziell unfairen und diskriminierenden Entscheidungen einer KI bzw. durch ML und ihrer Tragweite steigt entsprechend kontinuierlich an. Ziel dieses Beitrages war es daher, die verschiedenen potenziellen Verzerrungen im Bereich

des ML aufzudecken und diese in das Vorgehensmodell CRISP-DM nach Chapman et al. (2000) einzuordnen. Die Grundlage bildete eine strukturierte Literaturanalyse nach vom Brocke et al. (2015). Bei der systematischen Betrachtung der gesammelten Bias anhand des CRISP-DM Modells zeigt sich, dass in jeder Projektphase der Entwicklung Ursachen für Bias existieren. Grund dafür ist die direkte Beteiligung von Menschen in jedem Entwicklungsschritt eines ML-Modells. Den Phasen Datenverständnis und Datenaufbereitung des CRISP-DM Modells konnten die meisten Arten von Bias zugeordnet werden. Dementsprechend liegt der Fokus der meisten Arbeiten auf einem Bias, der sich in den gesammelten Daten befindet und in der Trainingsphase durch das ML-Modell übernommen wird (Cramer et al. 2018). Unsere Analyse der Ursachen für Bias im ML anhand des CRISP-DM Modells zeigt aber, dass ungeachtet dieser Fokussierung die Ursachen für Bias auch in anderen Entwicklungsschritten liegen können, da die menschliche Beteiligung vielfältig ist (Hundman et al. 2018).

Zusammenfassend bietet unser Beitrag somit einen Überblick über Verzerrungen durch menschliche Denkmuster und ihre Einflüsse auf den Entwicklungsprozess einer KI im weiteren Sinne bzw. eines ML-Modells im engeren Sinne. Allerdings ist unsere Forschung nicht ohne Limitationen. So bietet unsere strukturierte Literaturanalyse einen repräsentativen Überblick auf bekannte Bias aus forschungsbezogenen Publikationen. Je nach Anwendungsszenario kann es sein, dass sich weitere Bias ergeben. Dennoch kann dieser Beitrag die Grundlage für die Entwicklung fairer und ethischer KI-Entwicklungsprozesse und Reviews bilden. Gleichzeitig untermauert unser Beitrag die zukünftig steigende Relevanz der Forschung zu XAI, um potenzielle Bias innerhalb jeder Phase sichtbar zu machen und letztendlich dem Endanwender ein Werkszeug an die Hand zu geben, KI-Entscheidungen zu hinterfragen und nachzuvollziehen (Yan et al. 2020). Damit einhergehend zeigt sich jedoch auch, dass zukünftige Forschung diese Werkzeuge noch speziell für die Aufdeckung von Bias anpassen muss, um einen konfliktfreien Einsatz von KI in der Praxis zu ermöglichen.

Förderung Dieses Forschungs- und Entwicklungsprojekt wurde mit Mitteln des Bayerischen Staatsministeriums für Wirtschaft, Landesentwicklung und Energie (StMWi) innerhalb des Förderprogramms „Informations- und Kommunikationstechnik“ (Kennzeichen DIK0143/02) gefördert und vom Projektträger VDI + VDE Innovation + Technik GmbH betreut.

Funding Open Access funding enabled and organized by Projekt DEAL.

Literatur

vom Brocke J, Simons A, Riemer K, Niehaves B, Plattfaut R, Cleven A (2015) Standing on the shoulders of giants: challenges and recommendations of literature search in information systems research. CAIS 37(1):206–224

Chapman P, Clinton J, Kerber R, Khabaza T, Reinartz T, Shearer C, Wirth R (2000) CRISP-DM 1.0: Step-by-step data mining guide. https://www.kde.cs.uni-kassel.de/wp-content/uploads/lehre/ws2012-13/kdd/files/CRISPWP-0800.pdf. Zugegriffen: 8. Juli 2021

Cramer H, Garcia-Gathright J, Springer A, Reddy S (2018) Assessing and addressing algorithmic bias in practice. Interactions 25(6):58–63

Dobbe R, Dean S, Gilbert T, Kohli N (2018) A broader view on bias in automated decision-making: Reflecting on epistemology and dynamics. arXiv preprint. arXiv:1807.00553

Erfani S, Baktashmotlagh M, Moshtaghi M, Nguyen X, Leckie C, Bailey J, Kotagiri R (2016) Robust domain generalisation by enforcing distribution invariance. In: Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence New York, USA, S 1455–1461

Fjeld J, Achten N, Hilligoss H, Nagy A, Srikumar M (2020) Principled artificial intelligence: mapping consensus in ethical and rights-based approaches to principles for AI. Berkman Klein Center Research Publication (2020-1), S 1–39

Garcia-Gathright J, Springer A, Cramer H (2018) Assessing and addressing algorithmic bias-but before we get there. arXiv preprint. arXiv:1809.03332

Hanika H (2019) Digitalisierung, Big Data und Big To-dos aus Sicht der Rechtswissenschaft. In: Digitale Transformation von Dienstleistungen im Gesundheitswesen VI. Springer Gabler, Wiesbaden, S 67–87

Heindorf S, Scholten Y, Engels G, Potthast M (2019) Debiasing vandalism detection models at wikidata. In: The World Wide Web Conference San Francisco, USA, S 670–680

Hellström T, Dignum V, Bensch S (2020) Bias in Machine Learning—What is it Good for? arXiv preprint. arXiv:2004.00686

Henderson P, Sinha K, Angelard-Gontier N, Ke NR, Fried G, Lowe R, Pineau J (2018) Ethical challenges in data-driven dialogue systems. In: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society New Orleans, USA, S 123–129

Herm L-V, Wanner J, Seubert F, Janiesch C (2021) I don't get it, but it seems valid! The connection between explainability and comprehensibility in (X)AI research. In: European Conference on Information Systems. Virtual Conference, S 1–17

Holstein K, Wortman VJ, Daumé H III, Dudik M, Wallach H (2019) Improving fairness in machine learning systems: what do industry practitioners need? In: CHI Conference on Human Factors in Computing Systems Glasgow, Scotland, S 1–16

Hube C, Fetahu B, Gadiraju U (2019) Understanding and mitigating worker biases in the crowdsourced collection of subjective judgments. In: CHI Conference on Human Factors in Computing Systems Glasgow, Scotland, S 1–12

Hundman K, Gowda T, Kejriwal M, Boecking B (2018) Always lurking: understanding and mitigating bias in online human trafficking detection. In: Conference on AI, Ethics, and Society New Orleans, USA, S 137–143

Janiesch C, Zschech P, Heinrich K (2021) Machine learning and deep learning. Electron Mark 31:685–695

Jensen D, Neville J (2002) Linkage and autocorrelation cause feature selection bias in relational learning. In: Proceedings of the Nineteenth International Conference on Machine Learning ICML2002, San Francisco, USA, S 259–266

Li Y, Vasconcelos N (2019) Repair: removing representation bias by dataset resampling. In: Proceedings of the Conference on Computer Vision and Pattern Recognition Xi'an, China, S 9572–9581

Lutfi SL, Fernández-Martínez F, Lucas-Cuesta JM, López-Lebón L, Montero JM (2013) A satisfaction-based model for affect recognition from conversational features in spoken dialog systems. Speech Communication, 55(7):825–840

Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A (2021) A survey on bias and fairness in machine learning. ACM Comput Surv 54(6):1–35

Mujtaba DF, Mahapatra NR (2019) Ethical considerations in ai-based recruitment. In: International Symposium on Technology and Society (ISTAS) Boston, USA, S 1–7

Murphy KP (2012) Machine learning: a probabilistic perspective. MIT Press, Cambridge

Nier H (2018) Wie weiblich ist die IT? https://de.statista.com/infografik/13283/frauen-in-der-tech-branche/. Zugegriffen: 20. März 2021

Raab G, Unger A, Unger F (2010) Marktpsychologie. Springer, Wiesbaden

Sengupta E, Garg D, Choudhury T, Aggarwal A (2018) Techniques to eliminate human bias in machine learning. In: Proceedings of the International Conference on System Modeling & Advancement in Research Trends Moradabad, India, S 226–230

Sharma A, Wehrheim H (2019) Testing machine learning algorithms for balanced data usage. In: Conference on Software Testing, Validation and Verification (ICST) Xi'an, China, S 125–135

Slack D, Hilgard S, Jia E, Singh S, Lakkaraju H (2020) Fooling lime and shap: adversarial attacks on post hoc explanation methods. In: AAAI/ACM Conference on AI, Ethics, and Society New York, USA, S 180–186

Suresh H, Guttag JV (2019) A framework for understanding unintended consequences of machine learning. arXiv preprint. arXiv:1901.10002

Tang L, Liu H (2005) Bias analysis in text classification for highly skewed data. In: IEEE International Conference on Data Mining ICDM'05, Houston, USA, S 4–8

Udeshi S, Arora P, Chattopadhyay S (2018) Automated directed fairness testing. In: Proceedings of the 33rd ACM/IEEE International Conference on Automated Software Engineering New York, USA, S 98–108

Wanner J, Popp L, Fuchs K, Heinrich K, Herm L-V, Janiesch C (2021) Adoption barriers of AI: a context-specific acceptance model for industrial maintenance. In: European Conference on Information Systems. Virtual Conference

Wiemer H, Drowatzky L, Ihlenfeldt S (2019) Data mining methodology for engineering applications (DMME)—A holistic extension to the CRISP-DM model. Appl Sci 9(12):2407

Yan JN, Gu Z, Lin H, Rzeszotarski JM (2020) Silva: interactively assessing machine learning fairness using causality. In: Conference on Human Factors in Computing Systems. Virtual Conference, S 1–13

Zhang Q, Yang LT, Chen Z, Li P (2018) A survey on deep learning for big data. Inf Fusion 42:146–157