Synergy Consult

White Paper

Künstliche Intelligenz und Biases. Auch Maschinen haben Vorurteile

Warum KI verzerrte Ergebnisse auswirft und was Unternehmen dagegen tun können

Dr. Petra Köppel

Inhalt

Executive Summary

Dieses White Paper untersucht das Thema Künstliche Intelligenz (KI) und Biases (zu Deutsch: Verzerrungen) und bietet Unternehmen praktische Handlungsempfehlungen. Biases, die häufig unbewusst in KI-Systeme gelangen, können zu diskriminierenden oder schlicht falschen Ergebnissen führen. Angesichts der zunehmenden Nutzung von KI in Unternehmensprozessen – von der Personalauswahl über Kundeninteraktionen bis zur Entscheidungsfindung – stellt dies eine erhebliche Herausforderung dar.

Die Entstehung von Biases kann auf unterschiedlichen Ebenen stattfinden: durch System- und Entwicklerentscheidungen, durch die Auswahl und Strukturierung von Trainingsdaten sowie durch die Art der Nutzung. Verzerrungen beeinflussen nicht nur die Qualität der Ergebnisse, sondern stellen auch das Vertrauen der Nutzer_innen in die Technologie infrage.

Unser White Paper präsentiert ein Lösungsmodell auf vier Ebenen, um Biases in KI-Systemen gezielt entgegenzuwirken: Das Fundament bildet eine Unternehmenskultur, die auf Vielfalt und Einbeziehung diverser Perspektiven zielt. Durch klare Richtlinien und ethische Leitplanken im Rahmen einer Governance ist die Nutzung von KI zu steuern. Die Qualität und Repräsentativität der Trainingsdaten sind im Rahmen des Datenmanagements entscheidend für die Reduzierung von Biases. Schließlich führt der Kompetenzaufbau der Entwickler_innen und Anwender_innen zu einem verantwortungsvollen Umgang und stärkt die Rolle der Anwender_innen als Kontrollinstanz.

Mit diesem White Paper wollen wir Unternehmen unterstützen, den Herausforderungen rund um KI und Biases zu begegnen. Wir hoffen, dass das White Paper einen Beitrag zur Schaffung zugleich objektiver und fairer KI-Systeme liefert – damit Unternehmen nicht nur die Potenziale der KI ausschöpfen, sondern auch den Ansprüchen an Diversity, Equity & Inclusion (DEI) gerecht werden.

1 Warum wir uns bei Synergy Consult mit KI und Biases beschäftigen

Künstliche Intelligenz (KI) verändert die Arbeitswelt in rasantem Tempo und bietet Unternehmen erhebliche Potenziale, von Effizienzgewinnen bis hin zu datengetriebenen Entscheidungsprozessen. Gleichzeitig birgt der Einsatz von Künstlicher Intelligenz Risiken, insbesondere in Bezug auf die Wahrung von Diversity, Equity & Inclusion (DEI). Bei Synergy Consult sehen wir es als unsere Aufgabe, Unternehmen dabei zu unterstützen, diese Herausforderungen zu erkennen und ihnen zu begegnen. Ein besonderes Anliegen ist uns die systematische Einbindung der Fachexpert_innen aus Human Resources- und Diversity-Abteilungen, die in vielen Unternehmen bei der Etablierung von KI bislang noch nicht involviert sind. Diese Expert_innen können auf Basis ihrer bisherigen Erfahrung und Expertise jedoch zukünftig maßgeblich dazu beitragen, unbewusste Denkmuster zu identifizieren und zu vermeiden, die – unwissentlich in KI-Systeme implementiert – zu verzerrten Ergebnissen führen können, die sowohl benachteiligend wirken, aber auch falsche Unternehmensentscheidungen generieren.
Ein weiteres Problemfeld ist die fehlende klare Zuständigkeit für den Einsatz von KI innerhalb von Unternehmen. Dies führt oft zu einer individuellen, unsystematischen Nutzung, bei der grundlegende Aspekte wie Datenqualität, Transparenz und Fairness leicht übersehen werden. Zusätzlich zeigt sich, dass technische Kenntnisse allein für den sicheren und verantwortungsvollen Umgang mit KI nicht ausreichen. Vielmehr ist Artificial Intelligence Literacy, eine grundlegende Kompetenz in der Handhabung von KI-Technologie insbesondere auch zum Umgang mit Biases, bei allen Anwender_innen erforderlich.
Mit diesem White Paper möchten wir Wege aufzeigen, wie Unternehmen durch eine klare Verantwortungsstruktur, gezielte Schulungen und eine Kultur des bewussten Umgangs mit Technologie die Vorteile von Künstlicher Intelligenz nutzen und gleichzeitig den Ansprüchen von Diversity, Equity und Inclusion gerecht werden können.

2 Kontext und Problemlage: Künstliche Intelligenz und Biases

Generative KI, wie sie beispielsweise in Anwendungen wie ChatGPT oder Google Gemini eingesetzt wird, beschreibt Modelle, die mit Hilfe von maschinellem Lernen aus Trainingsdaten Muster erkennen und darauf basierend neue Inhalte generieren können. Diese Inhalte können Texte, Bilder, Audiodateien oder sogar Videos umfassen.
Zugrunde liegt eine große Menge an Daten, mit Hilfe derer das jeweilige KI-System Wahrscheinlichkeiten berechnet, um zu Ergebnissen zu kommen. Durch die umfassendere Datenbasis an personenbezogenen Daten können Algorithmen[1] gezielte Differenzierungen von Personen vornehmen.
Vor allem durch die Zusammenführung von Daten aus verschiedenen Quellen entstehen detaillierte Profile, die differenzierte Einteilungen in Kategorien erleichtern. Diese Wahrscheinlichkeitsschätzungen führen jedoch dazu, dass KI-Systeme - insbesondere Large Language Models (LLMs)[2] - auch ungewollte Verzerrungen (Biases) beinhalten.
Diese Biases können durch ungenaue Algorithmen, problematische Kategorisierungen oder fehlerhafte Annahmen in der Programmierung entstehen – insbesondere, wenn das zuständige Team weder divers besetzt noch für die Problematik sensibilisiert ist. Weitere Verzerrungen können über die Trainingsdaten einfließen – über einen Mangel an Repräsentation, historische Ungleichheiten, fehlerhafte Datenerhebungsmethoden oder die selektive Gewichtung bestimmter Datenarten. KI-Modelle bilden die gesellschaftlichen Werte und Strukturen ihrer Trainingsdaten nach, und wenn diese Daten von Biases beeinflusst sind, wird das Modell diese übernehmen.

[1] Ein Algorithmus ist eine präzise, schrittweise Anleitung zur Lösung eines Problems oder zur Durchführung einer bestimmten Aufgabe. Ein Algorithmus kann einfach sein, wie eine Rechenoperation, oder komplex, wie ein maschinelles Lernmodell, das aus großen Datenmengen Muster erkennt.

[2] Large Language Models (LLMs) sind KI-Modelle, die auf neuronalen Netzen basieren und mithilfe riesiger Textmengen trainiert wurden, um Sprache zu verstehen und zu generieren.

Ein solches KI-System ergibt dann im Ergebnis ein verzerrtes oder einseitiges Bild, welches unter Umständen kaum auf den ersten Blick zu erkennen ist. Die Konsequenz ist jedoch gravierend, da dies den Wahrheitsgehalt der KI-basierten Entscheidungen beeinflussen kann und zuweilen zu schlichtweg falschen Ergebnissen
führt. Zudem kann es dazu kommen, dass eine KI Menschengruppen diskriminiert.
Zusammengefasst treten Biases also in KI-Tools auf, die personenbezogene Daten verarbeiten, wie Sprachmodelle, Bild- und Gesichtserkennungssysteme, Empfehlungssysteme und automatisierte Entscheidungsprozesse im Personalwesen. Solche Anwendungen analysieren große Datenmengen, die menschliche Eigenschaften, Verhaltensweisen, Präferenzen oder demographische Daten abbilden. Diese Art der KI-Tools wird in allen Bereichen des Unternehmens genutzt.

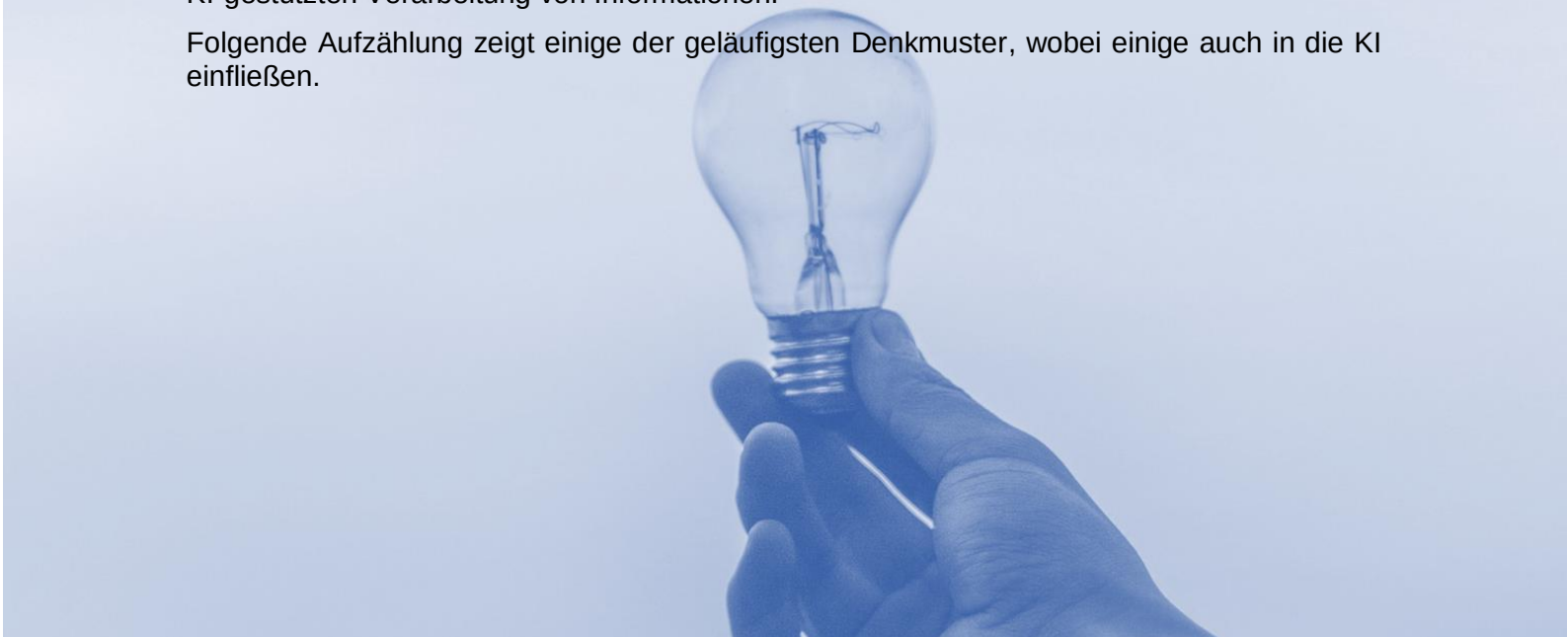
Sofort zu denken ist an die Ansprache und Gewinnung von Bewerber_innen, ans Talent Management oder alle anderen HR-Felder. Doch auch sämtliche Bereiche, die mit Kundendaten arbeiten, sind ähnlich betroffen – sei es das Marketing, der Vertrieb oder die Kundenverwaltung.
In unserem White Paper demonstrieren wir, auf welchem Weg genau Biases in KI-Systeme gelangen und vor allem, über welche Ebenen Unternehmen mit gezielten Maßnahmen diesem verzerrenden Einflussfaktor entgegentreten können, um den Umgang mit KI unternehmerisch sinnvoll und ethisch vertretbar zu gestalten.

Exkurs:
Biases oder unbewusste Denkmuster – eine psychologische Herleitung

Der Mensch ist darauf angewiesen, in hochkomplexen Umgebungen handlungsfähig zu bleiben. Indem das menschliche Gehirn eine Vielzahl von Informationen unbewusst und automatisch verarbeitet, hilft es, vor allem wiederkehrende Situationen schnell zu erkennen und angemessen darauf zu reagieren – dies wird ein unbewusstes Denkmuster genannt. Nur wenn eine Information besonders salient, also auffällig ist, schenkt der Mensch ihr explizite Aufmerksamkeit und bearbeitet sie aktiv – dies betrifft meist die schwierigen oder neuartigen Aufgaben.

Auf diese Weise teilen Menschen ihre Energie effektiv ein und agieren hoch funktional. So weit so gut. Das Problem liegt darin, dass in der automatischen Anwendung Muster auf Situationen angelegt werden, auf die sie nicht passen. Daher führen die unbewussten Denkmuster zuweilen zu falschen Bewertungen und Reaktionen. Dies ist der Grund, warum sie Verzerrungen, also Biases genannt werden. Dieser Begriff der Verzerrung wird nun auch verwendet in der KI-gestützten Verarbeitung von Informationen.

Folgende Aufzählung zeigt einige der geläufigsten Denkmuster, wobei einige auch in die KI einfließen.



Stereotype

ist ein Set an Merkmalen, die einer bestimmten sozialen Gruppe zugeschrieben werden und damit jedem einzelnen Mitglied dieser Gruppe. Ein Stereotyp ermöglicht schnelle Entscheidungen durch Vereinfachung und Kategorisierung und wird auch Vorurteil genannt. Stereotype sind zunächst neutrale Erwartungen und Vorstellungen darüber, wie sich Angehörige einer bestimmten sozialen Gruppe verhalten, wie sie aussehen oder welche Fähigkeiten sie haben könnten. Sie können zum Beispiel auf Alter, Geschlecht, Nationalität, Religion oder sexueller Orientierung beruhen und sind Teil des kollektiven Wissens einer Gesellschaft. Auch positive Verallgemeinerungen wie "Brillenträger sind klug", "Frauen sind einfühlsam" oder "Schwule sind kreativ" sind Stereotype.

Autoritätsgläubigkeit

ist die Tendenz, der Meinung einer Autoritätsperson (unabhängig von ihrem Inhalt) größeren Glauben zu schenken und sich davon maßgeblich beeinflussen zu lassen. Kritik und gegensätzliche Meinungen werden diesen Autoritäten gegenüber nicht geäußert. Autoritäten in Unternehmen sind natürlich die Vorgesetzten, aber auch Personen mit einer längeren Betriebszugehörigkeit oder einer stärkeren Persönlichkeit. Autoritätsgläubigkeit hat zur Folge, dass wichtige Erkenntnisse zurückgehalten werden und Entscheidungen ohne sie daher von geringerer Qualität sind. Dieser Effekt tritt besonders in hierarchischen Umgebungen auf.

Eigengruppenverzerrung

(auch Eigengruppenbevorzugung) ist die Tendenz, diejenigen zu bevorzugen, die der eigenen Gruppe angehören. Dies beruht auf dem Ähnlichkeitsdenken und fördert positive Beziehungen zu Menschen, die ähnliche Eigenschaften innehaben. Am Arbeitsplatz äußert sich dies darin, dass Personen Angehörige der eigenen Gruppe (der „Ingroup“) in ihrer Karriere unterstützen oder sie für besonders interessante Projekte auswählen. Diese Voreingenommenheit führt zu einer Ungleichbehandlung von Menschen am Arbeitsplatz. Die Tendenz, Mitglieder der Ingroup zu bevorzugen, kann dazu führen, dass die Outgroup benachteiligt wird.

Herdentrieb

ist die Neigung von Menschen (genau wie bei Tieren, deshalb der Begriff „Herde“), wie die Mehrheit zu handeln oder zu denken. Einzelne orientieren sich an dem, was das Kollektiv vorgibt. Menschen haben ein starkes Bedürfnis einer Gruppe anzugehören, denn das gibt Sicherheit und Zugehörigkeitsgefühl. So entwickeln sich oft Gruppennormen und -regeln, ohne dass sie offen diskutiert werden. Dies ist eine sehr starke Gruppendynamik und kann zu Gruppenkonformität und sogar -druck führen. Einzelne lassen sich ggfs. einschüchtern, sich zu äußern oder abweichende Informationen einzubringen. Oft fühlen sich Menschen, die anders denken, einen Anpassungsdruck, insbesondere wenn sie in der Minderheit sind.

Ähnlichkeitsdenken

ist ein unbewusstes Denkmuster (auch unter Mini-Me bekannt), bei dem eine Person danach beurteilt wird, wie ähnlich sie einem selbst ist. Personen neigen dazu, Menschen sympathischer zu finden, in denen sie sich selbst wiedererkennen. Diese Ähnlichkeit kann auf körperlichen Merkmalen beruhen, auf einer gemeinsamen Vergangenheit oder auf der Zugehörigkeit zur gleichen sozialen Gruppe. Dies kann zu einer homosozialen Reproduktion führen - dies bedeutet, dass sich ein System, das aus ähnlichen Menschen besteht, immer wieder nachbildet. Dieses Phänomen wird besonders bei Einstellungsentscheidungen deutlich. Es gilt als einer der Hauptgründe für den geringen Anteil von Frauen in Führungspositionen.

Bestätigungsfehler

ist die Tendenz, Informationen zu suchen, die mit den eigenen Überzeugungen übereinstimmen. Dies führt dazu, dass Menschen widersprüchliche Informationen ignorieren, so dass Entscheidungen nicht alle relevanten Sachverhalte berücksichtigen. Bestehende Überzeugungen formen die Erwartungen in einer bestimmten Situation und beinhalten Vorhersagen zu deren Ausgang. Die Wahrscheinlichkeit, dass Menschen bei ihren vorgefassten Meinungen bleiben, ist besonders hoch, wenn das Thema sehr wichtig und für sie selbst von Bedeutung ist. In einer stärkeren Ausprägung handeln Menschen sogar so, dass sie damit das erwartete Ergebnis herbeiführen; dies wird als selbsterfüllende Prophezeiung bezeichnet.

Status Quo Verzerrung

ist die Bevorzugung der aktuellen Situation gegenüber jeder Art von Veränderung. Der jetzige Stand wird als Referenzpunkt genommen, und jede Abweichung von dieser Basis wird als Gefahr empfunden. Selbst wenn neue Lösungen besser sind, zögern Menschen, vertraute Dinge aufzugeben. Für die meisten Menschen ist das Verlassen ihrer Komfortzone etwas Unangenehmes. Als Folge bleiben viele Menschen lieber beim Gewohnten und stehen Innovationen skeptisch gegenüber. Dies führt häufig dazu, dass sie Umstrukturierungsprojekte, neue IT-Tools oder eine größere Vielfalt innerhalb der Organisation ablehnen.

Vogel-Strauß-Taktik

beschreibt, wie Menschen häufig negative Informationen vermeiden. Anstatt sich mit der Situation auseinanderzusetzen, stecken die Menschen den Kopf wie Strauße den Kopf in den Sand, um der Gefahr nicht entgegenzublicken. Die Folge ist, dass ungünstige Fakten ignoriert werden. Manchmal denken die Menschen, dass dass ein Problem gelöst werden kann, indem man ihm keine Aufmerksamkeit schenkt. Unerwünschte Rück-meldungen oder Hinweise werden heruntergespielt oder bagatellisiert und nicht zur Verbesserung der Situation genutzt. Aber so zu tun, als gäbe es keine Probleme, kann die Dinge oft verschlimmern und zu Schäden führen, die verhindert worden wären, wenn man sich den Dingen direkt gestellt hätte.

3 Arten von Biases

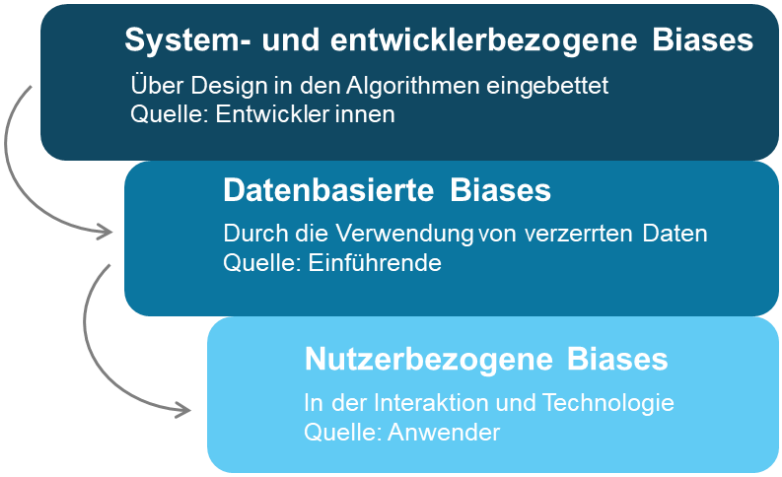


Abbildung: Drei Einfalltore für Biases in der KI

3.1 System- und entwicklerbezogene Biases

System- und entwicklerbezogene Biases entstehen, wenn Designentscheidungen und Werte der Entwickler_innen in die Entwicklung und Struktur von KI-Systemen einfließen. Diese Art von Bias ist tief in die Algorithmen und Funktionsweisen der KI eingebettet und kann während des Entwicklungsprozesses aufgrund subjektiver Perspektiven oder spezifischer Annahmen auftreten. Die Entwickler_innen entscheiden, welche Modelle, Parameter, Kategorien und Methodologien sie verwenden, was die Ergebnisse der KI beeinflusst. Wenn diese Entscheidungen durch kulturelle, sprachliche oder subgruppenbedingte Faktoren beeinflusst werden, können sie subtile oder sogar manifeste Verzerrungen in das System einbringen. Wenn wir einen Blick auf die Entwickler_innen werfen, fällt auf, dass sie oft männlich und weiß sind. Weiß deshalb, weil derzeit die USA das Feld der KI-Entwicklung (Tech-Unternehmen und KI-Startups wie Google, OpenAI, IBM, Microsoft, Meta und Amazon) dominieren. Auch europäische Länder wie Großbritannien, Deutschland und Frankreich sowie Kanada sind zu nennen – allesamt westliche Industrieländer mit deren Kultur und Werten.[3]

3.1.1 Algorithmic Bias

Der Algorithmusbias bezeichnet Verzerrungen, die durch das Design und die Struktur eines Algorithmus selbst verursacht werden. Diese Verzerrungen entstehen häufig durch bestimmte Annahmen und Gewichtungen, die während des Entwicklungsprozesses (meist unbewusst) getroffen werden. Wie Orwat in einer Studie der Antidiskriminierungsstelle des Bundes 2019 darlegt, werden im Bereich des Data-Mining[4] und maschinellen Lernens Algorithmen so gestaltet, dass sie bestimmte Zielgrößen vorhersagen oder schätzen, die den Zweck der Analyse definieren. Diese Ziele müssen jedoch in berechenbare Funktionen umgewandelt werden, die anhand historischer oder aktueller Daten optimiert werden. Ein Beispiel ist die Bestimmung der Kreditwürdigkeit von Antragstellenden oder der Leistungskraft von Mitarbeitenden. Entwickler_innen müssen entscheiden, was „gute Kreditwürdigkeit“ oder „gute Leistung“ bedeutet und wie sie gemessen werden.

[3] China ist zwar neben USA ebenso ein großer Akteur (Baidu, Alibaba und Tencent sowie staatlich unterstützte Programme), aber dessen KI-Tools kommen in Europa weniger zum Zuge.

[4] Data Mining ist der Prozess, bei dem große Datenmengen analysiert werden, um Muster, Zusammenhänge und nützliche Informationen zu entdecken.

Diese Zielfunktionen, in denen bestimmte Variablen minimiert oder maximiert werden, bergen jedoch Diskriminierungsrisiken. Zum Beispiel können Auswahlprozesse, die auf bestimmten Zielvariablen wie Betriebszugehörigkeit basieren, Gruppen benachteiligen, die häufiger den Arbeitgeber wechseln (wie Frauen), obwohl ihre Leistung vergleichbar ist.
Ein weiteres Risiko besteht in der Auswahl und Definition von Kategorien (zu Englisch: Labels). Da Labels die Grundlage der Entscheidungslogik im Algorithmus bilden, beeinflussen sie direkt, wie Menschen in neuen Datensätzen klassifiziert werden. Sind diese Kategorien jedoch subjektiv gewählt oder unklar definiert, tragen sie oft bestehende Vorurteile in den Algorithmus hinein. Zum Beispiel kann ein Label wie „passender Kandidat“ auf persönlichen Vorstellungen beruhen und diskriminierende Annahmen widerspiegeln, die Gruppen aufgrund bestimmter Merkmale ausschließen.
Zudem kann die begrenzte Auswahl der Einflussvariablen (Features) im Modelltraining Risiken bergen. Die Reduzierung auf grobe Merkmale kann dazu führen, dass die Modelle nicht genügend differenzieren und dadurch höhere Fehlerquoten bei bestimmten Gruppen auftreten.

3.1.2 Framing Bias

Der Framing Bias oder -Effekt beschreibt, wie die Präsentation von Informationen die Entscheidungsfindung beeinflusst. Wie Vinney in ihrem Artikel 2024 darstellt, werden Nutzer_innen in ihren Entscheidungen nach der Art der Darstellung von Informationen beeinflusst. Sie zeigt, dass positive oder negative Formulierungen zu unterschiedlichen Reaktionen führen können, selbst wenn der Inhalt identisch ist. Dies ist besonders relevant für KI-Systeme, die Informationen ohne Kontext oder Alternativen präsentieren. Dies mag den Eindruck erwecken, es gäbe nur eine relevante Lösung.

3.1.3 Ethical Bias

Der ethische Bias bezieht sich auf die moralischen Vorstellungen, die unbewusst oder bewusst in die Entwicklung von KI-Systemen einfließen. Entwickler_innen treffen in der KI-Erstellung verschiedenste Entscheidungen, die auf ihren persönlichen Werten basieren. Dies kann dazu führen, dass KI-Systeme systematisch bestehende Ungleichheiten verstärken.
Eine Studie von Noble (2018) beleuchtet die Problematik des Ethical Bias in KI, insbesondere in Suchmaschinen. Sie zeigt, dass Suchalgorithmen rassistische und sexistische Stereotype reproduzieren können. Algorithmen, die vermeintlich neutral gestaltet sind, liefern durch die Auswahl und Gewichtung der Trainingsdaten durch die Entwickler_innen ethisch problematische Ergebnisse.
Verschärft wird diese Tendenz durch den Bias Blind Spot (Verzerrungsblindheit) – die Tendenz, das eigene Wahrnehmen, Denken, Erinnern und Urteilen für objektiv und neutral zu halten. Wenn dementsprechend Entwickler_innen von KI-Systemen nicht in der Lage sind, ihre eigenen Biases zu erkennen, führt dies zu einem unkritischen Vertrauen in die von ihnen geschaffenen Systeme und die Verweigerung, einen darin enthaltenen ethischen Bias zu erkennen.

3.2 Datenbasierte Biases

Datenbasierte Biases beziehen sich auf Verzerrungen, die durch die Qualität, Herkunft und Auswahl der Daten entstehen, die zur Schulung von KI-Systemen verwendet werden. Die Trainingsdaten, auf denen KI-Algorithmen basieren, spiegeln häufig die Verzerrungen wider, die in den zugrunde liegenden Informationen enthalten sind. Wenn beispielsweise historische Daten verwendet werden, die gesellschaftliche Ungleichheiten widerspiegeln, übernimmt die KI diese und führt zu Diskriminierungen. Diese Biases können zudem durch unrepräsentative oder unvollständige Datensätze verstärkt werden, was zu einer fehlerhaften Entscheidungsfindung führt. Die Genauigkeit und Ausgewogenheit der Ergebnisse der KI hängt daher stark davon ab, wie sorgfältig die Daten ausgewählt, gefiltert und bereinigt werden.

3.2.1 Selection Bias

Der Selection Bias tritt auf, wenn die Auswahl der Trainingsdaten für ein KI-Modell nicht repräsentativ für die Zielpopulation ist. Diese Verzerrung entsteht, wenn bestimmte Gruppen oder Merkmale in den Daten über- oder unterrepräsentiert sind, was bewirkt, dass das Modell bestimmte

Ergebnisse bevorzugt oder andere vernachlässigt. Der Selection Bias beeinflusst die Generalisierbarkeit des Modells und führt oft dazu, dass es für bestimmte Benutzergruppen oder Szenarien nicht zuverlässig arbeitet.
Der Selection Bias kann aus einer Reihe von Ursachen resultieren:
Datenbeschaffung: Wenn die Daten aus einer begrenzten Quelle stammen, kann dies eine Überrepräsentation bestimmter Merkmale zur Folge haben.
Methoden zur Datensammlung: Wenn Daten aus Umfragen oder bestimmten Kanälen gesammelt werden, die nur bestimmte Bevölkerungsgruppen anziehen.
Bewusste oder unbewusste Auswahl: Wenn bestimmte Daten bewusst oder unbewusst ausgeschlossen werden, z. B. Daten aus einem bestimmten Zeitraum oder aus bestimmten geografischen Regionen.
Buolamwini & Gebru (2018) fanden heraus, dass kommerzielle Gesichtserkennungssoftware je nach Geschlecht und Hautfarbe unterschiedlich genau arbeiten. Die Autorinnen zeigten, dass die Klassifizierungsgenauigkeit bei weißen Männern um bis zu 99% liegt, während sie bei schwarzen Frauen nur bei 65% liegt.

3.2.2 Representation Bias

Der Representation Bias tritt zutage, wenn die zugrunde liegenden Daten nicht ausreichend die Diversität oder Komplexität der realen Welt abbilden. Hier liegt der Fokus auf der ungleichen oder unvollständigen Abbildung bestimmter Gruppen, Kategorien oder Merkmale in den Daten. Der Representation Bias beeinträchtigt genau wie der Selection Bias die Fähigkeit des KI-Modells, faire und ausgewogene Entscheidungen zu treffen, da es bestimmte Gruppen bevorzugt oder andere diskriminiert.

Abgrenzung zwischen Selection Bias und Representation Bias
Selection Bias bezieht sich auf die Nicht-Repräsentativität der Auswahl insgesamt, d. h. die Art und Weise, wie die Daten gesammelt werden, und ob diese Daten überhaupt die Zielpopulation repräsentieren.
Representation Bias hingegen bezieht sich darauf, wie umfassend und gleichmäßig die Diversität innerhalb der ausgewählten Daten abgebildet ist, also ob alle Gruppen oder Merkmale gleichermaßen und in angemessener Tiefe vertreten sind.
In einem Satz zusammengefasst: Selection Bias beschreibt das Problem der Datenquellen und Auswahlmethoden, während Representation Bias die fehlende oder verzerrte Darstellung von bestimmten Gruppen oder Eigenschaften innerhalb der Daten beschreibt. Beide Bias-Arten führen jedoch dazu, dass KI-Modelle fehlerhafte oder unfaire Entscheidungen treffen, besonders wenn sie auf eine diverse Bevölkerung angewendet werden.

3.2.3 Cultural Bias

Der kulturelle Bias bezieht sich auf die Verzerrungen, die auftreten, wenn Künstliche Intelligenz (KI) auf Daten trainiert werden, die Normen und Werte einer Kultur widerspiegeln. Diese Daten sind Texte aus Büchern, Websites und sozialen Medien. Diese stammen jedoch zumeist aus dem Land und der Sprache, in dem das KI-Tool entwickelt wurde, und enthalten natürlicherweise die dortigen Werte. Zum Beispiel sind viele Trainingsdaten stark auf englischsprachige Inhalte fokussiert, die typischerweise Werte und Normen aus westlichen, individualistischen Kulturen präsentieren. Während des Trainingsprozesses lernt das Modell, Muster in den Daten zu erkennen und darauf basierende Vorhersagen zu treffen. In der Konsequenz sind die Modelle nicht in der Lage, kontextuelle Nuancen oder kulturelle Sensibilitäten zu erkennen; sie sind darauf optimiert, die häufigsten und wahrscheinlichen Ausdrücke aus ihren Trainingsdaten zu verwenden.
Zudem generieren sie kulturell geprägte Antworten: Yan Tao et al. (2024) untersuchten, wie kulturelle Vorurteile in den Antworten von fünf weit verbreiteten großen Sprachmodellen (GPT-4, 4-turbo, 4o, 3.5-turbo und 3) präsent sind. Die Ergebnisse zeigen, dass alle Modelle kulturelle Werte aufweisen, die vorwiegend englischsprachigen und protestantischen europäischen Ländern ähneln. Die KI neigt dazu, Antworten zu produzieren, welche individuelle Selbstverwirklichung widerspiegeln, wie beispielsweise Individualismus, persönliche Freiheit und Selbstentfaltung. Diese Tendenz ist typischerweise in westlichen Ländern zu finden, insbesondere in den USA und Westeuropa. Dadurch entsteht eine Überrepräsentation westlicher

Werte in den KI-generierten Inhalten. Das hat zur Konsequenz, dass Perspektiven und Werte aus anderen Kulturen, die weniger stark auf individuelle Selbstverwirklichung fokussiert sind, möglicherweise nicht ausreichend repräsentiert oder sogar ignoriert werden.

3.2.4 Historic Bias

Der historische Bias beschreibt die Tendenz von KI-Systemen, vergangene gesellschaftliche Ungleichheiten und Vorurteile in ihren Entscheidungen und Prognosen zu reproduzieren. Dies geschieht, weil viele KI-Modelle auf historischen Daten basieren, die bereits bestehende Diskriminierungen widerspiegeln. Solche Systeme geben dann zuweilen Antworten aus, die in der Gegenwart inzwischen als problematisch identifiziert werden.
Ein typisches Beispiel findet sich im Bereich der Personalauswahl, wo KI-Systeme zur Analyse von Bewerberdaten eingesetzt werden. Wenn ein KI-Algorithmus auf einem Datensatz trainiert wird, der in der Vergangenheit geschlechtsspezifische Ungleichheiten aufwies, wird er wahrscheinlich diese unterschiedliche Repräsentanz lernen und fortführen.

Folgender Fall zeigt die Problematik: Dastin beschreibt 2018 in der New York Times, wie Amazon ein KI-gestütztes Rekrutierungstool entwickelte, das Bewerbungen analysierte und die besten Kandidat_innen auswählte. Bei der Auswertung der Bewerberdaten stellte sich heraus, dass das System männliche Bewerber gegenüber weiblichen bevorzugte, basierend auf historischen Daten, in denen die Mehrheit der erfolgreichen Bewerber Männer waren. Als die Fachleute des Unternehmens versuchten, das System zu optimieren, wurde deutlich, dass die KI diskriminierende Muster lernte, indem sie von den bestehenden Daten, die bereits geschlechtsspezifische Vorurteile beinhalteten, beeinflusst wurde. Amazon stellte schließlich die Verwendung des KI-Systems ein, weil es nicht in der Lage war, geschlechtergerechte Entscheidungen zu treffen, und die bestehenden Vorurteile verstärkte.

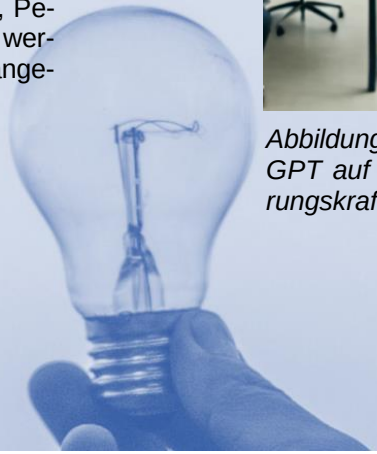
Exkurs:

Ein Experiment mit ChatGPT zum historischen Bias

Wir haben bei ChatGPT folgenden Prompt eingegeben: „Erstelle mir ein Bild einer Führungskraft eines mittelständischen Unternehmens“. Dieses finden Sie an dieser Stelle abgebildet.

Es zeigt deutlich den historischen Bias: In der Vergangenheit war die Mehrheit deutscher Führungskräfte männlich, weiß, zwischen 35-50 Jahre alt und üblicherweise im Anzug zu sehen. Daher gibt ChatGPT dies selbst in 2024 aus und trägt diesen historischen Fakt fort. Frauen, People of Colour, Menschen mit Behinderung werden von der KI nicht als Führungskräfte angegeben.

Abbildung: Die Ausgabe von ChatGPT auf den Prompt zu einer Führungskraft



3.3 Nutzerbezogene Biases

Nutzerbezogene Biases entstehen durch die Interaktionen von Menschen mit KI-Systemen und die Art und Weise, wie Entscheidungen und Technologien von ihnen wahrgenommen werden. Die Menschen bringen ihre Annahmen in die Nutzung der Technologie ein, was die Ergebnisse der KI beeinflussen kann. Nutzerbezogene Biases können sich darauf auswirken, wie die KI reagiert, und dabei unbewusste oder bewusste Vorlieben oder Abneigungen der Nutzer_innen widerspiegeln. Darüber hinaus entstehen weitere Verzerrungen, wenn die Anwender_innen Ergebnisse falsch interpretieren, überbewerten oder nicht kritisch prüfen.

3.3.1 Automation Bias

Der Automation Bias bezieht sich auf die Tendenz von Menschen, Entscheidungen und Empfehlungen von Künstlicher Intelligenz (KI) übermäßig zu vertrauen, selbst wenn diese möglicherweise fehlerhaft sind. Benutzer_innen neigen dazu, den Fähigkeiten und Vorschlägen von KI-Systemen zu vertrauen, da sie oft als objektiv und rational wahrgenommen werden.
In der Studie von Lee et al. (2019) wird untersucht, wie Menschen in verschiedenen Situationen auf automatisierte Systeme reagieren und wie ihr Vertrauen in diese Systeme das Verhalten beeinflusst. Die Forscher fanden heraus, dass User_innen oft dazu neigen, Entscheidungen von KI-Systemen zu überbewerten, was zu einer verminderten kritischen Auseinandersetzung mit den Ergebnissen führt. Insbesondere wird die Tendenz dokumentiert, Entscheidungen der KI in sicherheitskritischen Kontexten, wie etwa in der Luftfahrt oder im Gesundheitswesen, als unverzichtbar zu betrachten, auch wenn die Systeme Fehler aufweisen oder ungenaue Informationen liefern. Das Problem wird nach den MIT-Forschern um Peter Park dadurch verschärft, dass Sprachmodelle wie GPT-4 von OpenAI inzwischen in der Lage sind, sehr überzeugend zu argumentieren und auch auf Täuschungen und Lügen auszuweichen.

3.3.2 Confirmation Bias

Der Confirmation Bias oder Bestätigungsfehler beschreibt die Tendenz von Nutzer_innen, Informationen zu bevorzugen, die ihre bestehenden Überzeugungen unterstützen, während sie gegenteilige Beweise ignorieren oder abwerten. In der Interaktion mit KI-Systemen kann dieser Bias gravierende Auswirkungen auf die Entscheidungsfindung haben. Er kann dazu führen, dass wichtige alternative Perspektiven übersehen werden, wodurch die Qualität der Entscheidungen leidet.
Besonders problematisch ist dies in der Forschung, wo KI-Systeme zur Datenanalyse und Ergebnisinterpretation eingesetzt werden. Eine aktuelle Studie von Bashkirova und Krpan (2024) untersucht den Bestätigungsfehler in der Interaktion zwischen Fachleuten im Bereich der psychischen Gesundheit und KI-Systemen. Die Forscher_innen fanden heraus, dass Psycholog_innen und Psychologiestudierende dazu neigen, KI-Empfehlungen zu vertrauen und diese anzunehmen, wenn sie mit ihren eigenen Diagnosen übereinstimmen. Besonders auffällig war, dass erfahrenere Fachleute skeptischer gegenüber KI-Vorschlägen waren, wenn diese von ihren eigenen Einschätzungen abwichen, was den Einfluss des Bestätigungsfehlers verstärkt.

4 Unser Lösungsmodell auf vier Ebenen

Die Integration von Künstlicher Intelligenz (KI) in Unternehmen stellt nicht nur eine technische, sondern auch eine kulturelle und organisatorische Herausforderung dar. Der Erfolg einer KI-Implementierung hängt wesentlich davon ab, wie gut die Belegschaft diesen Veränderungsprozess versteht, unterstützt und darin eingebunden wird. Eine gezielte Begleitung des Change-Prozesses ist unerlässlich, um sicherzustellen, dass alle Mitarbeitenden die Notwendigkeit und den Mehrwert der Neuerung erkennen und bereit sind, diese zu akzeptieren und anzuwenden.
Ein erfolgreicher Change-Prozess beginnt damit, dass die Mitarbeitenden über die Ziele und Vorteile der KI-Implementierung informiert werden. Der Sinn und konkrete Nutzen von KI sollten klar kommuniziert werden, um Akzeptanz zu fördern und das Verständnis für KI als unterstützendes Werkzeug zu schaffen.
Häufig bestehen Unsicherheiten, die mit der Einführung von KI einhergehen – sei es die Sorge, dass KI den Menschen ersetzen könnte, oder die Angst vor einer unkontrollierten Verselbständigung der Technologie. Indem diese Probleme proaktiv thematisiert und sachlich entkräftet werden, lassen sich potenzielle Widerstände frühzeitig mindern.
Ein weiteres Hindernis im Change-Prozess kann Desinteresse unter Mitarbeitenden sein, die sich nicht direkt mit der Neuerung auseinandersetzen möchten. Und bei denjenigen, die sich über den Einzug von KI freuen, ist häufig ein Nichtwissen über die Relevanz von Bias zu beobachten. Um dem vorzubeugen, ist die Schaffung eines Problembewusstseins entscheidend.
Der Change-Prozess zielt somit darauf ab, eine Kultur des Vertrauens und des bewussten Umgangs mit Künstlicher Intelligenz zu schaffen, in der Mitarbeitende die Technologie als Unterstützung begreifen und sich des Einflusses von Biases bewusst sind. Eine solche Veränderung stellt sicher, dass KI erfolgreich eingeführt wird und die gesamte Organisation von den Vorteilen eines kritischen und ethischen Umgangs mit KI profitieren kann.
Wir möchten im Folgenden vier Ebenen skizzieren, auf denen für einen solch verantwortungsvollen Umgang mit KI gesorgt werden kann:

4.1 Ebene Unternehmenskultur

Unternehmenskultur dient als Fundament, das die Werte eines Unternehmens widerspiegelt und eine entscheidende Rolle dabei spielt, wie Mitarbeitenden mit KI-basierten Technologien umgehen. Der Umgang mit Biases in KI beginnt hier. Folgendes sollte bei der diversity-bewussten Einführung von KI berücksichtigt werden:

1. Erwartungsmanagement hinsichtlich KI betreiben

An allererster Stelle ist Klarheit zu schaffen: Unternehmen sollten transparent machen, warum und wie KI eingesetzt wird, und eine klare Zielvorgabe definieren, die das spezifische Problem aufzeigt, das die KI lösen soll. Damit wird vermieden, dass unrealistische Erwartungen an KI-Tools entstehen, und die Mitarbeitenden verstehen den Sinn und Zweck der KI-Anwendungen besser.

2. Dialog über bias-freie KI anstoßen

Die Organisation sollte eine bewusste Entscheidung darüber treffen, welche ethischen Grundsätze sie im Umgang mit KI verfolgt, und diese klar formulieren. Dies sollte sowohl in Zusammenhang mit der Unternehmensstrategie als auch mit den Unternehmenswerten angegangen werden, sodass Kriterien wie Transparenz, Fairness und Inklusion systematisch als Grundlage etabliert werden. Mit dieser Basis kann dann ein entsprechender Dialog in der Belegschaft angestoßen werden.
Diversity Manager_innen sind die Brückenbauer: Sie übernehmen eine Schlüsselrolle – diese beginnt damit, indem sie diesen Dialog zu ethischer KI im gesamten Unternehmen initiieren und darin für die Bias-Thematik sensibilisieren. Die Diskussionen sollten sich am Stand und den Bedürfnissen der verschiedenen Zielgruppen orientieren – die Botschaften sollten für jede_n verständlich und relevant sein.
In diesem Zusammenhang können die Diversity-Verantwortlichen zusammen mit Menschen mit Inklusionsbedarf das Thema Barrierefreiheit von KI aufgreifen – denn digitale Medien bieten einen enormen Mehrwert, um Menschen mit Behinderungen noch besser zu inkludieren, wenn sie entsprechend gestaltet werden.

3. Bewusstsein für Problematik auf allen Ebenen schaffen

Um Transparenz zu KI und Biases zu schaffen, sollte parallel an verschiedenen Kanälen gearbeitet werden – im Folgenden ein paar Beispiele, um möglichst alle Mitarbeitergruppen aufzuklären:
Die Entwicklung und Verwendung von kurzen Videos können die Herausforderungen und Risiken, die mit Biases in KI einhergehen, anschaulich machen. Dies fördert ein grundlegendes Verständnis bei den Mitarbeitenden und kann als leicht zugängliches Format auf internen Plattformen genutzt werden.
Im Intranet können die verschiedenen Communities dazu angeregt werden, KI und Biases aus ihrer Sicht zu beleuchten.
Ein effektiver Weg, Bewusstsein zu schaffen, besteht darin, praxisbezogene Anwendungsbeispiele aufzuzeigen, z. B. anhand von Stellenanzeigen, die gezielt Diversität ansprechen und Bias vermeiden. Solche Beispiele veranschaulichen, wie Bias in der Praxis reduziert werden kann.

4. Eigenverantwortlicher Umgang als Ziel

Im Endeffekt hängt das Ergebnis einer KI-Ausgabe immer vom jeweiligen Endnutzer ab. Daher ist bei dem Dialog in den Fokus zu nehmen, dass jede_r Einzelne eine entsprechende Eigenverantwortung entwickelt.
Dies beginnt mit Data Knowledge: Mitarbeitende sollten verstehen, wie die Daten verarbeitet werden und welche Entscheidungsmechanismen im System ablaufen. Dies umfasst ein grundlegendes Wissen über Datenstrukturen und den Einfluss, den die Mitarbeitenden selbst auf die Ergebnisse der KI haben, indem sie ihre eigenen Daten teilen (Data Sharing) und das System fortwährend nutzen und gleichzeitig trainieren (Maintenance). Damit wird klar, dass die Verantwortung für die Pflege und die Qualität der KI von allen gleichzeitig zu tragen ist.

5. Fehlerkultur etablieren

Die Unternehmenskultur sollte eine offene Fehlerkultur unterstützen, die Mitarbeitenden erlaubt, neue Ansätze auszuprobieren, zu lernen und dabei selbst besser zu werden. Gleichzeitig werden dann auch die

Prozesse optimiert und KI-Lösungen verbessert, auch im Hinblick auf Bias-Reduktion.

6. Diverse Teams in allen relevanten Bereichen

Von diesem Punkt geht die meiste Wirkung aus. Der positive Einfluss von Diversität in Teams auf Problemlösekompetenz ist wissenschaftlich mehrfach bestätigt worden (z.B. bei Stahl et al. 2010) und sollte auch im vorliegenden Einsatzbereich beachtet werden. Die Besetzung von KI-Entwicklungsteams und Change-Management-Gruppen mit diversen Mitarbeitenden stellt sicher, dass verschiedene Perspektiven in die KI-Entwicklung und -Implementierung einfließen, so dass Biases bereits in ihrer Entstehung minimiert werden. Über ein inklusives Miteinander, also die freie Meinungsäußerung der Betroffenen (seien es Frauen, People of Colour, Ältere und sämtliche andere Diversity-Gruppen), wird das Problem sofort erkannt und kann gebannt werden. Doch auch die Teams, die jeweils zuständig sind für das KI-Training, dem Testen oder für Schulungen sind divers zu besetzen, um an den verschiedenen Stadien mögliche Verzerrungen wahrzunehmen und entgegenzusteuern.

Dieser kulturelle Ansatz hilft Unternehmen nicht nur, Biases in KI zu vermeiden, sondern auch ein Umfeld zu schaffen, in dem Transparenz, Eigenverantwortung und kontinuierliches Lernen im Umgang mit KI gefördert werden.

4.2 Ebene Governance

Eine erfolgreiche und verantwortungsbewusste Einführung von KI in Unternehmen erfordert eine klare Governance-Struktur, die ethische Leitplanken setzt und die relevanten Stakeholder einbindet. Die Governance-Ebene spielt eine entscheidende Rolle dabei, den Rahmen für eine bias-freie, faire und transparente Nutzung von KI zu schaffen.

1. Teil der Strategie: Top-Down-Verankerung durch Unternehmensleitung

Eine erfolgreiche Governance beginnt an der Spitze des Unternehmens: Die Geschäftsführung bzw. Vorstand sollten das Thema KI und Ethik (inkl. Bias) aktiv in die Unternehmensstrategie aufnehmen. Diese Verankerung im Leitbild signalisiert der gesamten Belegschaft und den Stakeholdern, dass der ethische Umgang mit KI eine Priorität ist und dass Maßnahmen zur Bias-Reduzierung und Fairness einen hohen Stellenwert haben.

2. Einbindung der DEI-Verantwortlichen

Eine zentrale Maßnahme zur Förderung einer ethischen KI-Nutzung ist die systematische Einbindung von Diversity, Equity & Inclusion (DEI)-Verantwortlichen in alle KI-Prozesse. DEI-Expert_innen können dabei helfen, potenzielle Diskriminierungen zu identifizieren und bias-bezogene Fragestellungen aufzuwerfen. Ihre Mitwirkung im KI-Steuerungsteam oder in einem Gremium für ethische KI stellt sicher, dass das Thema Diversität vom Moment der KI-Einführung kontinuierlich im Fokus bleibt und dass ein aktiver Dialog über Herausforderungen und Lösungen angestoßen wird. Die für DEI verantwortliche Person hat die Verantwortung, in den unterschiedlichen Phasen der KI-Einführung, - Training und Anwendung auf die spezifischen Bias-Probleme hinzuweisen und idealerweise anzuregen, den vorliegenden Maßnahmenkatalog zur Umsetzung zu bringen.

3. Gremium für ethische KI

Es ist dringend zu empfehlen, ein Gremium für ethische KI einzuführen. Es steuert und kontrolliert die KI-Nutzung und stellt sicher, dass diese in Übereinstimmung mit den ethischen Standards des Unternehmens erfolgt. Zusammengesetzt aus Vertreter_innen der Bereiche HR, Legal, IT, Datenschutz und wie soeben dringend empfohlen aus DEI, stellt es sicher, dass ethische, rechtliche und technische Anforderungen erfüllt werden. Dieses Gremium setzt Prozesse auf, die eine bias-freie Einführung von KI sicherstellen. Es prüft neue Anwendungsfälle, indem es potenzielle KI-Einsatzbereiche bewertet und ethische Risiken identifiziert. Es stößt Schulungsprogramme für alle oben genannten Zielgruppen an. Es überwacht Audits, Bias-Tests und Modellupdates und greift ein, wenn Auffälligkeiten festgestellt werden (Kontrollprozess s.u.). Bei unvorhergesehenen Herausforderungen kann das Gremium zum Trouble Shooting angerufen werden.

Darüber hinaus sollten weitere Anlaufstellen und Meldestellen eingerichtet werden. Neben dem Betriebsrat, der beispielsweise eine_n Referent_in für KI benennen kann, könnten Datenschutz, IT und HR als Ansprechpartner_innen für Fragen und Anliegen im Zusammenhang mit KI agieren. Auch anonyme Meldekanäle wie eine Speak-up-Hotline oder ein Whistleblowing-System können genutzt werden, um Missstände oder potenzielle Biases zu melden.

4. Ethische Richtlinien und Betriebsvereinbarung

Die Entwicklung ethischer Richtlinien, die als Leitfaden für die Nutzung von KI dienen, ist ein weiterer wichtiger Baustein der Governance. Diese schaffen klare Grundsätze für die Entwicklung und Anwendung von KI im Unternehmen. Da nicht alle Eventualitäten durch Prozesse abgedeckt werden können, geben ethische Grundsätze in einem Code of Conduct den Mitarbeitenden Orientierung und stärken das Verantwortungsbewusstsein. Eine ergänzende Betriebsvereinbarung stellt sicher, dass alle Mitarbeitenden einheitliche Vorgaben haben.

5. Kontrollprozesse

Es sollten Prozesse eingeführt werden, die kontinuierlich überwachen, ob KI-Systeme fair und transparent arbeiten. Diese Maßnahmen sind entscheidend, um Biases frühzeitig zu erkennen und das Vertrauen in die KI-gestützten Entscheidungsprozesse aufrechtzuerhalten.
Dies beginnt bereits bei der Anschaffung, wo bereits eine detaillierte Evaluierung durchgeführt werden sollte. Diese sollte auch ein Augenmerk auf mögliche Biases in den Algorithmen legen und prüfen, wie bias-frei die Systeme laufen. Ein wichtiger Teil der Prüfung ist die Sicherstellung der Konformität mit dem Allgemeinen Gleichbehandlungsgesetz (AGG).
Des Weiteren sind regelmäßige Audits und Tests anzusetzen, die halbjährlich durchgeführt und bei Modellanpassungen oder Datenupdates wiederholt werden, um Biases zu verhindern. KPI-basierte Kontrollmechanismen können Fehlerraten in Bezug auf unterschiedliche Gruppen oder die Anzahl positiver und negativer Fehlentscheidungen abbilden und bei Abweichungen ein rechtzeitiges Eingreifen ermöglichen. Die kontinuierliche Verbesserung wird durch Protokolle gesichert, die aufzeichnen, welche Probleme identifiziert und wie sie behoben wurden. Klare Verantwortlichkeiten und regelmäßige Berichte zu den Kontrollergebnissen schaffen Transparenz und fördern das Vertrauen der Belegschaft in die KI-Prozesse.

6. Externe Expertise

Zusätzlich zur internen Überwachung ist es sinnvoll, externe Expert_innen für KI und Biaes einzubeziehen, um insbesondere in kleineren Organisationen die entsprechende Expertise in den Betrieb zu holen. Zudem kann damit eine unabhängige, objektive Sichtweise gewonnen werden, um mögliche Schwächen in den Prozessen zu identifizieren, sollte es keine interne Kontrollinstanz geben.

Zusammenfassend schafft eine solide Governance-Struktur die Voraussetzung für einen verantwortungsvollen Einsatz von KI und gewährleistet, dass Biases frühzeitig erkannt und adressiert werden. Die gezielte Integration ethischer Grundsätze und die interdisziplinäre Zusammenarbeit sind entscheidend, um KI als fairen und zuverlässigen Bestandteil der Unternehmensstrategie zu etablieren.

4.3 Ebene Daten

4.3.1 Datenmanagement

Die Verantwortung für die Minimierung von Biases in den Daten liegt sowohl bei den Unternehmen, die die KI entwickeln, als auch bei den Abnehmern, die diese Systeme nutzen. Jedes KI-Tool wird durch die KI-Anbieter vortrainiert, doch in vielen Fällen gibt es die Möglichkeit, dass im Anschluss die erwerbenden Unternehmen ein Fine-Tuning mit eigenen Daten vornehmen. Im Datenmanagement ist nämlich auf folgende Aspekte zu achten:

1. Diversität der Datensätze

Ein vielfältiger und umfassender Datensatz, mit dem das KI-System trainiert wird, der eine breite Palette von Merkmalen wie Geschlecht, ethnische Herkunft, Alter und sozioökonomischen Status abdeckt, kann helfen, den historischen Bias, den kulturellen Bias, den Representation Bias oder Selection Bias zu reduzieren. So werden möglichst viele Facetten der realen Welt berücksichtigt, um repräsentative Ergebnisse zu erzielen.

2. Datenaufbereitung und Vorverarbeitung

Vor dem Training eines Modells sollten die Daten gründlich auf potenzielle Verzerrungen analysiert werden. Dies umfasst das Entfernen oder Angleichen von Kategorien, die über- oder unterrepräsentiert sind, und die Berücksichtigung eines Gleichgewichts zwischen verschiedenen Gruppen.

3. Bewusste Datenkennzeichnung

Datenkennzeichnung in der KI bezeichnet den Prozess, bei dem Rohdaten mit spezifischen Labels oder Annotationen versehen werden, um diese für das Training und die Evaluierung von maschinellen Lernmodellen nutzbar zu machen. Bei der Kennzeichnung von Daten durch Menschen sollte auf potenzielle Biases der Durchführenden geachtet werden. Ein divers zusammengesetztes Team, das die Daten kennzeichnet, kann helfen, die Vielfalt an Perspektiven zu wahren und Verzerrungen zu verringern.

4. Ausgleich der Repräsentanz

Wenn bestimmte Gruppen oder Szenarien im Datensatz unterrepräsentiert sind, kann die Generierung synthetischer Daten helfen, diese Lücken zu schließen. So lässt sich sicherstellen, dass alle relevanten Gruppen ausreichend repräsentiert sind, ohne auf reale Daten zugreifen zu müssen, die schwer oder gar nicht zu beschaffen sind. Zudem können Daten regularisiert gewichtet werden – das bedeutet, dass die Modelle so trainiert werden, dass sie verschiedene Gruppen mit angepassten Gewichtungen behandeln, um sicherzustellen, dass Gruppen mit einer geringeren Anzahl an Daten nicht benachteiligt werden. Dies wird explizit gemacht, um den historischen Bias, den Repräsentations- oder Selektionsbias auszuschalten. Allerdings führt dies in manchen Fällen zu Folgefehlern, indem bei Maßgabe einer kulturell gemischten Darstellung von Menschengruppen auf Bildern dann auch Menschen zu sehen sind, die in diesem Kontext in der Realität nicht zu finden sind.

5. Gezielte Datenauswahl

So viel Daten wie nötig, so wenig wie möglich – nach diesem Prinzip sollten Unternehmen sich darauf konzentrieren, nur die für ihre spezifischen Anwendungsfälle relevanten Daten zu verwenden. Dies hilft, Überfrachtung mit unnötigen Informationen zu vermeiden, die potenziell Biases einführen können. Denn nicht die Quantität, sondern die Qualität ist entscheidend: Hochwertige Daten, die gut kuratiert und verifiziert sind, minimieren die Wahrscheinlichkeit von Verzerrungen und verbessern die Leistung des KI-Modells. Dazu gehört die Überprüfung der Daten auf Genauigkeit, Relevanz und Konsistenz.

6. Kontinuierliches Monitoring und Auditing der Daten

Ein wiederkehrendes Monitoring der Daten nach dem Training ist sinnvoll, um festzustellen, ob neue Daten das Modell in eine verzerrte Richtung lenken. Regelmäßige Audits und Analysen der verwendeten Daten können helfen, schleichende Biases frühzeitig zu erkennen und zu beheben.
Eine regelmäßige Bereinigung der Daten sollte ebenso durchgeführt werden, um fehlerhafte, irrelevante oder veraltete Daten zu entfernen. Dies kann durch automatisierte Tools oder manuelle Überprüfungen geschehen.

7. Verwendung von Bias Detection Tools

Anwendungen, welche die KI auf Biases überprüft, sollten als Standard etabliert werden. Ein Bias Detection Tool testet die Eingabedaten auf potenzielle Biases, analysiert die Verteilung von Merkmalen wie Geschlecht und Ethnizität und identifiziert Über- oder Unterrepräsentationen. Darüber hinaus bewertet das Tool die Leistung von KI-Modellen auf verschiedenen Datensätzen und verwendet Gleichheitsmetriken wie Demographic Parity, um festzustellen, ob Vorhersagen für unterschiedliche Gruppen gleich sind. Das Tool identifiziert auch Quellen von Biases, wie Datenfehler oder unausgewogene Merkmale im Modell, und bietet Strategien zur Minderung von Biases. Schließlich generieren Bias Detection Tools Berichte, die gefundene Biases und empfohlene Maßnahmen dokumentieren, um Transparenz und Compliance zu gewährleisten.

8. Transparente Dokumentation

Es ist wichtig, die Datenquellen, -verarbeitungen und die angewandten Algorithmen klar zu dokumentieren. Dies erhöht die Transparenz und ermöglicht eine bessere Nachvollziehbarkeit der Entscheidungen des KI-Systems.

4.3.2 Interaktion mit Anwender_innen

1. Antwortstrategien

In Fällen von diskriminierenden Prompts können KI-Systeme so programmiert werden, dass sie neutrale oder allgemeine Antworten geben, die keine Diskriminierung unterstützen. Die Systeme können zudem darauf ausgelegt werden, auf diskriminierende Eingaben hinzuweisen, um das Bewusstsein für die Problematik zu schärfen. KI-System sollten keinerlei diskriminierende Prompts annehmen.
Zusätzlich können sie regelbasierte Ansätze verwenden, um bestimmte diskriminierende Begriffe oder Phrasen zu erkennen und entsprechend zu reagieren, indem sie diese blockieren oder melden. Indem Nutzungsrichtlinien erstellt werden, können Nutzer_innen bei schwerwiegenden Verstößen gesperrt werden.

2. Benutzerfeedback

Benutzerfeedback ist entscheidend für die Verbesserung von KI-Systemen im Umgang mit diskriminierenden Inhalten. Durch einfache, intuitive Meldemöglichkeiten können Benutzer_innen problematische Inhalte schnell und anonym melden. Eine klare Kategorisierung des Feedbacks hilft, Muster zu erkennen. Echtzeit-Reaktionen des Systems auf Rückmeldungen fördern die Interaktivität. Die aggregierten Daten ermöglichen es Entwickler_innen, Trends zu identifizieren und das KI-Modell entsprechend anzupassen. Regelmäßige Berichterstattung über gemeldete Vorfälle schafft Transparenz. Schulungen und Community-Beteiligung sensibilisieren die User_innen für die Bedeutung des Feedbacks und tragen zur Schaffung einer respektvollen Umgebung bei.

3. Nutzung von Personas

KI-Inhalte können durch den Einsatz von Personas geprüft werden. Diese künstlichen Personas werden nach unterschiedlichen demografischen Merkmale erstellt, um verschiedene Diversity-Gruppen abzubilden. Indem diese Personas in den Entscheidungsprozess integriert werden, kann das System gezielt auf potenzielle Diskriminierungen aus Sicht dieser Personengruppen hin überprüft werden: Die KI kann so programmiert werden, dass sie die Reaktionen dieser Personas auf verschiedene Inputs simuliert, wodurch Schwächen im System sichtbar gemacht werden. Regelmäßige Überprüfungen des Outputs im Kontext dieser Personas helfen, die Fairness der Ergebnisse zu gewährleisten. Zudem fördert dieser Ansatz ein tieferes Verständnis für die Auswirkungen von Entscheidungen auf unterschiedliche Gruppen.

4. Erläuterungen zum Output

Die KI erklärt den User_innen, welche Informationen und Kriterien zur Erstellung eines Ergebnisses verwendet wurden. Bei der Generierung eines Bildes zum Beispiel legt die KI schriftlich dar, welche Eingaben, Parameter und Datenquellen sie berücksichtigt hat. Diese Transparenz ermöglicht es den Benutzer_innen, die Entscheidungsprozesse der KI besser zu verstehen. Indem die KI explizit macht, welche Interpretationen und Annahmen sie getroffen

hat, wird die Qualität des Outputs nachvollziehbarer. Dies fördert nicht nur das Vertrauen in die Technologie, sondern hilft auch, potenzielle Biases zu erkennen. Diese Technik unterstützt auch eine dialogorientierte Interaktion, in der Benutzer_innen gezielte Anpassungen vornehmen können.

4.4 Ebene Kompetenz

Datenmanagement allein reicht nicht aus, um das Problem von Biases in KI zu lösen; das Bewusstsein, die Fähigkeiten und das Wissen aller beteiligten Fachkräfte als Bestandteil von AI Literacy sind maßgeblich, um verantwortungsbewusst mit KI-Technologien umzugehen und korrigierend einzugreifen.

1. Zielgruppen

Zunächst benötigen Programmierende eine umfassende Schulung, um bei der Entwicklung von KI-Systemen auf potenzielle Biases in den Algorithmen und Datenquellen zu achten. Ebenso ist die Schulung der Tester_innen entscheidend, da sie sicherstellen müssen, dass die KI-Modelle unter realistischen Bedingungen auf Verzerrungen überprüft werden. Support-Teams sollten ebenfalls geschult werden, um Anfragen und Rückmeldungen zu Biases effektiv zu bearbeiten und Nutzerprobleme zu verstehen. Schließlich ist die Schulung für Anwender_innen von großer Wichtigkeit, damit sie die KI verantwortungsbewusst nutzen und die Grenzen sowie die Auswirkungen von Biases in den Ergebnissen verstehen. In letzter Instanz fungieren die Anwender_innen als Kontrollinstanz, denn es sollte in den vorangegangen Abschnitten deutlich geworden sein, dass die Problematik von Biases technisch nicht ausgeschlossen werden kann.

2. Inhalte zur Schulung von Anwender_innen als Kontrollinstanz

Zunächst ist es wichtig, eine Einführung in die Künstliche Intelligenz zu geben, die grundlegende Konzepte erklärt - wie KI funktioniert und wo sie eingesetzt wird. Dazu gehört auch ein Überblick über unterschiedliche Arten von KI, wie maschinelles Lernen und neuronale Netze. Ebenso sollte verdeutlicht werden, was KI in der Lage ist zu leisten und welches Tool für welche Zwecke geeignet ist.

Ein zentrales Thema sollte das Verständnis von Bias sein, einschließlich einer Definition der verschiedenen Arten von Bias, sowie der Ursachen und Auswirkungen von Bias in KI-Systemen – so wie in diesem White Paper ausführlich dargestellt.
Der zentrale Aspekt ist die verantwortungsvolle Nutzung von KI, und hierbei steht die eigene kritische Auseinandersetzung und Prüfung der KI-Outputs im Vordergrund. Denn die KI kann einerseits Diversity unterrepräsentieren (z.B. aufgrund des historischen Biases), aber auch überrepräsentieren (wenn sie gewichtet trainiert wurde).
Zudem ist Sensibilität zu schaffen, dass jede_r Nutzer_in gleichzeitig Trainer_in ist. Über Prompting und das Teilen von Daten in unternehmensinternen KI formen die Nutzer_innen die KI – daher müssen sie sich ihrer eigenen Biases im Klaren sein, um diese nicht an die KI weiterzugeben.

3. Formate für Anwenderschulungen

Präsenzschulungen bieten die Möglichkeit für gemeinsames Ausprobieren, interaktive Diskussionen und unmittelbares Feedback von Trainer_innen. Webinare ermöglichen eine breitere Reichweite und flexible Teilnahmezeiten. E-Learning-Kurse bieten den Vorteil, dass Teilnehmer_innen in ihrem eigenen Tempo lernen und parallel Erlerntes sofort in die Praxis umsetzen können.
Auf jeden Fall sollten reale Anwendungen eingesetzt werden, um Prompts und Ausgaben analysieren zu lassen und ein besseres Verständnis für die Auswirkungen von Bias zu entwickeln. Zu empfehlen ist, KI-generierte Outputs (vor allem Abbildungen von Menschen) zu zeigen, denn nichts sensibilisiert so sehr wie verzerrte Bilder zu sehen. Denn auf solchen ist zu erkennen, dass die KI oft nicht nur diskriminiert, sondern auch objektiv falsche Ergebnisse auswirft.
Wir empfehlen, die üblichen technischen Einführungsschulungen aus Diversity-Sicht zu ergänzen, um synergetisch und frühzeitig zu wirken. Ein spielerischer Zugang und Learning by Doing helfen, die Bias-Problematik zu verstehen und aktiv eigene Lösungsstrategien zu entwickeln. Über Gamification kann der ggfs. als anstrengend empfundene Lernprozess attraktiv, interaktiv und damit nachhaltiger gestaltet werden.
Sollten Schulungen nicht möglich sein, ist auf jeden Fall ein Leitfaden zu erstellen, in

dem die Unternehmenswerte nochmal aufgegriffen, ethische Fragestellungen erläutert und schließlich der kompetente Umgang mit KI schrittweise dargelegt werden. Ein solcher Leitfaden geht speziell auf ein KI-Tool ein und wird idealerweise für eine Zielgruppe zugschnitten – z.B. Recruiter_innen, Kundenbetreuer_innen, etc.

4. Richtiges Prompting

Für eine vorurteilsfreie und inklusive Nutzung von KI-Modellen ist das richtige Prompting entscheidend. Hier sind einige Empfehlungen für Anwender_innen:
Vorurteilsfreies Prompting: Prompts sollten möglichst neutral formuliert werden ohne die Verwendung von Stereotypen. Ein neutraler Sprachstil reduziert das Risiko, dass die KI durch voreingenommene Anfragen beeinträchtigt wird und führt zu objektiveren Ergebnissen.
Inklusive Sprache: Inklusive und gendergerechte Sprache im Prompting (z. B. durch Gendern) wirkt sich oft direkt auf die Antworten aus - so lassen sich bias-freie Ergebnisse in der Ausgabe erzielen.
Cleveres Prompting: Die Anwender_innen sollen das KI-Modell aktiv zur Bias-Reduktion auffordern. Durch Anweisungen wie „Berücksichtige verschiedene Perspektiven“ oder „Liefere eine kulturübergreifende Antwort“ kann das Modell bewusst in Richtung weniger voreingenommener Antworten gelenkt werden.
Kulturelles Prompting: Tao et al. bewiesen in ihren Studien, dass wenn Anwender_innen bei ChatGPT explizit angeben, aus welchem kulturellen Kontext die Antwort stammen soll, Biases maßgeblich reduziert werden. Eine Anpassung des kulturellen Hintergrunds im Prompting kann die kulturelle Relevanz und Neutralität der Antwort erhöhen – die Autor_innen nennen dies kulturelles Prompting.

5 Fazit

Das vorliegende White Paper zu KI und Biases zeigt auf, wie tiefgreifend die Auswirkungen von Verzerrungen in der Künstlicher Intelligenz auf die Qualität der Arbeit im Unternehmen sind.

Um diesen Herausforderungen zu begegnen, ist ein dezidiertes Verständnis der Problematik grundlegend, welches als Teil der Einführung von KI im Betrieb mit aufgebaut werden muss. Eine diverse und inklusive Unternehmenskultur hilft dabei. Eine Governance-Struktur für ethische KI liefert Orientierung und Prozesse für Einführung und Nutzung. Zentral ist der verantwortungsvolle Umgang mit Daten, um die dort verankerten Biases möglichst weitgehend zu minimieren. Doch letztendlich bleiben die einzelnen Anwender_innen die letzte Kontrollinstanz – entscheidend ist also deren Kompetenz.

Durch die Umsetzung dieser vier Lösungsebenen können Unternehmen sicherstellen, dass KI-Anwendungen fairer, transparenter und objektiver gestaltet werden – und somit das volle Potenzial der Technologie ohne diskriminierende Verzerrungen ausgeschöpft wird. Ein zukunftsfähiger Umgang mit KI erfordert daher kontinuierliche Reflexion, Weiterentwicklung und verantwortungsbewusstes Handeln auf allen Ebenen.

Danksagung

Die Grundlagen für dieses White Paper stammen aus der Synergiewerkstatt #41 ‚KI und Biases. Warum der nächste CEO Thomas heißt' des Netzwerks ‚Synergie durch Vielfalt' powered by Synergy Consult.

Das Netzwerk ‚Synergie durch Vielfalt' ist eine Initiative, die sich dem Ziel verschrieben hat, durch Wissens- und Erfahrungsaustausch die Potenziale der Vielfalt zu fördern und Unternehmen bei der Umsetzung von Diversity, Equity und Inclusion zu unterstützen. Die Synergiewerkstätten wie diese bringen Fachleute unterschiedlicher Disziplinen zusammen, um gemeinsam zukunftsweisende Lösungen zu entwickeln.

Gastgeber unserer diesmaligen Synergiewerkstatt war metafinanz Informationssysteme GmbH, die mit Geschäftsführer Rainer Göttmann und Staff Development Manager Thorsten Kolwe die Plattform geschaffen haben, auf der sich Fachleute mit den Herausforderungen von Biases in KI auseinandersetzen konnten. An dieser Stelle möchte ich ihnen sowie dem gesamten Team bei der metafinanz herzlich danken für die Gastfreundschaft, die wertvollen Inputs und das gemeinsame Engagement.

Unsere Gäste waren Diversity-Verantwortliche verschiedener deutscher und internationaler Unternehmen sowie KI-Expert_innen, an die ich ebenso ein großes Dankeschön aussprechen möchte. Durch ihre aktive Beteiligung, ihre unterschiedlichen Perspektiven und ihre Expertise ist es gelungen, tiefgehende Lösungsansätze zu entwickeln, welche die Basis des White Papers bilden.

Autorin

Dr. Petra Köppel ist Gründerin und Leiterin von Synergy Consult, einem Beratungsunternehmen, das auf die strategische Umsetzung von Diversity, Equity und Inclusion (DEI) in Unternehmen spezialisiert ist. Mit einem akademischen Hintergrund in Personal und Organisation sowie über 15 Jahren Erfahrung in der Beratung unterstützt sie Unternehmen dabei, eine wertschätzende und inklusive Kultur aufzubauen.

Synergy Consult

Unser Name ist Programm: Über Diversity, Equity und Inclusion – also über Vielfalt, Chancengleichheit und konstruktives Miteinander – entstehen Synergien. Da ergibt das Gesamte mehr als die Summe der Einzelteile. Vielfalt steht für einen Business Case und sorgt für:

- eine attraktive Arbeitgebermarke,
- motivierte und engagierte Mitarbeitende,
- ein zukunftsfähiges und innovatives Unternehmen,
- Kundenorientierung und Marktnähe,
- unternehmerischen Erfolg.

Synergy Consult bietet Consulting & Trainings von A wie Audit bis Z wie Zielkontrolle. Gemeinsam mit dem Vorstand entwickelt Dr. Petra Köppel und ihr Team eine fürs Unternehmen passende DEI-Strategie, anschließend schulen sie Führungskräfte zu Inclusive Leadership, liefern Werkzeuge für diverse Teams und sorgen für Partizipation aller Mitarbeitergruppen. E-Tools zur Diversity-Kompetenzentwicklung und das Netzwerk ‚Synergie durch Vielfalt' als Austauschplattform für Diversity-Profis komplettieren das Portfolio.

Um bei der aktuell zunehmenden Einführung von KI-Tools in Unternehmen Biases im Moment des Entstehens zu minimieren, liefert Synergy Consult Vorträge und Schulungen für die verschiedensten Zielgruppen. Diese reichen von einer generellen Sensibilisierung zur Thematik bis hin zu konkreten Anwenderschulungen.

Weitere Informationen über Synergy Consult finden Sie unter www.synergyconsult.de.

Quellen

Bashkirova, A., & Krpan, D. (2024). Confirmation bias in AI-assisted decision-making: AI triage recommendations congruent with expert judgments increase psychologist trust and recommendation acceptance. Computers in Human Behavior: Artificial Humans, 2(1).

Buolamwini, J., & Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. Proceedings of Machine Learning Research, 81, 77–91.

Vinney, C. (2024). The Framing Effect: How Perception Shapes Decision-Making. Verywell Mind.

Lee, J. D., & See, K. A. (2019). Trust in Automation: Designing for Appropriate Reliance. Human Factors, 61(1), 1-16.

Noble, S. U. (2018). Algorithms of Oppression: How Search Engines Reinforce Racism. New York University Press.

Orwat, C. (2019). Diskriminierungsrisiken durch Verwendung von Algorithmen.

Park, P. S., Goldstein, S., O'Gara, A., Chen, M., & Hendrycks, D. (2024). AI deception: A survey of examples, risks, and potential solutions. Patterns, 5(5).

Stahl, G. K., Maznevski, M. L., Voigt, A., & Jonsen, K. (2010). Unraveling the Effects of Cultural Diversity in Teams: A Meta-Analysis of Research on Multicultural Work Groups. Journal of International Business Studies, 41(4), 690–709.

Suresh, H., & Guttag, J. V. (2021). A Framework for Understanding Unintended Consequences of Machine Learning. Communications of the ACM, 64(2), 62–71.

Tao, Y., Viberg, O., Baker, R. S., & Kizilcec, R. F. (2024). Cultural Bias and Cultural Alignment of Large Language Models. PNAS Nexus, 3(9).

November 2024

Synergy Consult
Nymphenburger Str. 124a
D-80636 München
Tel.: +49 / 8106 / 211 62-88
info@synergyconsult.de