Fairness aus Prinzip[1]

Reflexionsleitfaden für faire KI-Entscheidungen in Unternehmen

Anna-Sophia Christ

1. Der Anspruch faire (KI-)Entscheidungen

Die fortschreitende Entwicklung von KI hat einen breit angelegten Diskurs über Fairness in unternehmerischen Entscheidungsprozessen angestoßen. Unternehmen stehen zunehmend in der Verantwortung, Fairness als grundlegendes Prinzip in die Entwicklung und Implementierung von KI zu integrieren (vgl. John-Mathews et al. 2022; Zhang 2024). Diese, bislang weitgehend selbstregulierte, Praxis orientiert sich oftmals an etablierten Standards wie dem IEEE P7003TM „Algorithmic Bias Considerations" oder der ISO/IEC RT 240027:2021 „Bias in KI-Systemen und KI-gestützter Entscheidungsfindung" (vgl. ISO/IEC 2021). Jüngst wurde diese Entwicklung in der EU durch die gesetzlichen Vorgaben des AI Acts weiter gefestigt, der spezifische Anforderungen an die Überprüfung von Diskriminierung stellt. Die konkrete Umsetzung dieser Vorgaben bleibt jedoch den Unternehmen überlassen und stellt sie vor die Herausforderung, tief verwurzelte und teils unbekannte Verzerrungen in historischen Daten sowie die Komplexität der Erklärbarkeit von fairen KI-Entscheidungen zu bewältigen (vgl. Alvarez et al. 2024; Kattnig et al. 2024). Weitere regulatorische Präzisierung wird aktuell im Rahmen des Digital Fairness Acts diskutiert. Ob und wie weitreichend dieser Unternehmen Vorgaben setzten wird ist jedoch noch offen, ein Gesetzesentwurf wird im zweiten Halbjahr 2026 erwartet.

Damit reiht sich der Anspruch an Fairness in eine grundsätzliche Herausforderung der digitalen verantwortlichen Unternehmensführung ein: den Prinzipien-Praxis-Transfer (vgl. Christ 2025; Mittelstadt 2019; Sanderson et al. 2023; Schiff et al. 2021). Der Prinzipien-Praxis-Transfer

[1] Bei der Erstellung des Kapitels wurden KI-Tools spezifisch und abgegrenzt zur Verbesserung des Lesefluss eingesetzt.

im Kontext der Corporate Digital Responsibility (CDR) zielt darauf ab, eine grundlegende ethische Fragestellung von Prinzipien – Wie wollen wir uns verhalten? – mit der praktischen Umsetzung und Steuerung zu verbinden – Wie erreichen und erhalten wir dieses Verhalten? (vgl. Christ et al. 2024; Lobschat et al. 2021; Mueller 2022). Dies lässt sich übertragen auf das Prinzip KI-Fairness in Unternehmensentscheidungsprozessen, bei dem verantwortungsvolle KI-Entwicklung und KI-Anwendung als Anspruch gesetzt wird. In Abhängigkeit von Stakeholdern und Kontext gibt es nicht ein einheitliches Verständnis und eine Metrik von Fairness (vgl. Carey/Wu 2023). Das Kapitel orientiert sich am Verständnis der CDR-Initiative: „Prinzip 6: Fairness – Wir stärken bei der Entwicklung und dem Einsatz technischer Systeme die Teilhabe und den Zugang zu einer digitalisierten Welt. Verbraucherinnen und Verbraucher werden nicht diskriminiert" (2025). Zur Präzision bei „nicht diskriminiert" handelt es sich um „die Abwesenheit von Vorurteilen oder Bevorzugung einer Person oder einer Gruppe aufgrund ihrer angeborenen oder erworbenen Eigenschaften" (Mehrabi et al. 2021: 2), die ohne, dass sie sachbezogen sind, zu strukturellem Ausschluss führen.

Zahlreiche bestehende Ansätze, insbesondere aus technischer Perspektive oder mit Fokus auf die Arbeit mit Data Scientists, setzen sich mit der Herausforderung von Verzerrungen in Daten auseinander. Verzerrungen können zu fehlender Fairness führen, mit Folgen für Betroffene und das Ergebnis, und stellt eine der zentralen Herausforderungen für datengetriebene Entscheidungsprozesse dar (vgl. Friedman/Nissenbaum 1996). Die Forschung im Bereich KI-Fairness fokussiert sich dabei meist auf technisch-mathematische Lösungsansätze (z. B.: Alvarez et al. 2024; Pant et al. 2025). In diesem Kapitel hingegen wird der Fokus auf eine steuernde unternehmensstrategische Ebene gelegt, denn bislang findet die praktische Anwendung entsprechender Prinzipien auf Unternehmensebene nur begrenzt statt (vgl. Beutel et al. 2019; Madaio et al. 2022).

Für eine wirksame Unternehmenssteuerung ist entscheidend die Etablierung einer vorangestellten reflektierenden Leitlinie, die den gesamten Prozess der Datenerhebung und -auswertung strukturiert und steuert, da dieser als potenzielle Hauptquelle für Fairnessdefizite betrachtet wird. Dabei liegt der Analyse die vereinfachte Annahme zugrunde, dass das Digitale lediglich eine Abbildung des Analogen ist (vgl. Nassehi 2019). Dies impliziert, dass eine kritische Reflexion der analogen Informationen, die als Grundlage für KI dienen, erforderlich ist, um mit der Abbildung von fehlender Fairness im Digitalen umzugehen (vgl. Kamiran/Calders 2009). Um die Herausforderungen, die Unternehmen im Hinblick auf KI-Fairness in Entscheidungsprozessen erfahren

adäquat zu adressieren, wird im Folgenden zunächst das zugrunde liegende Verständnis von Entscheidungslogik dargelegt. Darauf aufbauend wird eine Reflexionsleitfaden ausgeführt, der es Unternehmen ermöglicht, die Prämissen ihrer Entscheidungsfindung zu hinterfragen und neue Wege für faire KI-Entscheidungen zu evaluieren. Es wird explizit von KI-Entscheidungen gesprochen bei denen Menschen aktiv eingebunden sind (Human-in-the-Loop) oder überwachend eingreifen können (Human-on-the-Loop), da dies derzeit die häufigsten Anwendungsszenarien sind. Eine dritte Variante ist die vollständig automatisierter Entscheidungsfindung, ohne Zutun von Menschen (Human-out-of-the-Loop) (vgl. Crootof et al. 2022). Am Beispiel Bewerbungsprozess differenziert: Erstens, die KI liefert nur eine Vorbewertung. Zweitens, die KI interagiert automatisiert, der Mensch hat aber Einblick und kann intervenieren. Drittens, der Prozess läuft durch die KI automatisiert ab, ohne dass ein Mensch eine Bewerbung prüft. Konzeptionell zentral ist dabei die Einflussnahme von Menschen auf die Entwicklung und Anwendung von KI und Verantwortungsübernahme für Entscheidungen, die letztendlich wieder Menschen basierend auf KI treffen.

2. Analoge Entscheidungslogik übertragen auf KI

KI durchdringt Unternehmensprozesse und kann zunehmend Aufgaben der Entscheidungsfindung übernehmen. Sei es im Recruiting, wo Algorithmen Lebensläufe auf Eignung analysieren, im Einkauf zur Optimierung von Lieferketten, bei der Vergabe von Krediten, der Berechnung von Versicherungstarifen oder zur Unterstützung strategischer Marktanalysen. Diese Entwicklungen bedienen die unternehmerischen Anforderungen an Effizienz- und Funktionalitätsgewinn hervorragend (vgl. Krakowski et al. 2023). Entscheidungen, die Menschen stundenlanges Abwägen und wiederholtes Strukturieren von Informationen abverlangen, können von KI in Bruchteilen der Zeit durchgeführt werden. Zudem unterliegen Entscheidungsprozesse von Menschen dem variierenden Risiko der Subjektivität und des menschlichen Versagens. Mitarbeitende in der Personalabteilung, die hunderte Lebensläufe durchsehen, können unmöglich alle Kriterien in perfektem Vergleich behalten und völlig objektiv entscheiden. Ähnlich verhält es sich in der Kreditvergabe: Wer jahrelang erfahren hat, dass ein bestimmter Indikator mit einer höheren Auswahlwahrscheinlichkeit korreliert, wird diesen Faktor intuitiv stärker gewichten, meist ohne sich umfangreich rechtfertigen zu müssen, es wird Bestandteil beruflicher Praxis.

Die kognitiven Prozesse, die der unternehmerischen Entscheidungsfindung zugrunde liegen, lassen sich anschaulich durch die Arbeiten des Wirtschaftsnobelpreisträgers Daniel Kahneman (2012) erklären, der die menschliche Kognition in zwei Systeme unterteilte: das schnelle, intuitive System 1 und das langsame, reflektierende System 2. Im wiederkehrenden Tagesgeschäft verlassen sich Entscheidungstragende oftmals auf System 1. Diese Form des Denkens basiert auf implizitem Wissen und Erfahrungswerten – sie ermöglichen Expert*innen auf bekannte Muster zurückzugreifen und Entscheidungen effizient sowie scheinbar mühelos zu treffen, ohne dabei umfassende Informationsverarbeitung betreiben zu müssen. Diese Fähigkeit zur schnellen und sicheren Entscheidungsfindung gilt nicht nur als Kennzeichen von Expertise, sondern auch als essenzielle Voraussetzung, um der schieren Menge an Entscheidungen, die in Unternehmen täglich getroffen werden müssen, überhaupt gerecht zu werden. So kann auch die Entscheidungsfindung des Personalers und die der Bankangestellten im oben dargelegten Beispiel erklärbar gemacht werden.

Gleichzeitig birgt die Dominanz von schnellen, reaktiven Routineentscheidungen gewisse Risiken, insbesondere im Hinblick auf kognitive Verzerrungen, die durch die Anwendung heuristischer Strategie entstehen und in Abwägung mit den ressourcenintensiven reflektierenden Entscheidungen stehen. Hier setzt die Diskussion über den Einsatz von KI als Unterstützung oder gar Substitut, menschlicher Entscheidungsprozesse an. KI ist prinzipiell in der Lage, in sehr kurzer Zeit eine Vielzahl relevanter Informationen zu verarbeiten, wodurch sie eine Art „Turbo-System 2" darstellen könnte – ein rationales, datenbasiertes Korrektiv zum impulsiven System 1. Auf der Seite des erhöhten Ressourceneinsatzes steht dabei nur noch Rechenleistung und Daten sowie der damit einhergehende Energieverbrauch. Jedoch zeigte uns die Einführung von DeepSeek im ersten Halbjahr 2025, dass bisher angenommene Aufwände reduzierbar sind. Die chinesische open-source KI-Chatbot Plattform schaffte es ihr Modell mit wesentlich weniger Daten zu trainieren und den Rechenaufwand bei Abfragen durch begrenzten zugeschnittenen Einbezug von Informationen im Vergleich zu anderen Anbietern wie OpenAI zu reduzieren.

Aus hegemonietheoretischer Perspektive bedarf einer kritischen Reflexion, wann KI tatsächlich neutral oder objektiv sind. Denn auch algorithmische Entscheidungsprozesse beruhen auf Heuristiken – auf vereinfachenden Suchstrategien, die zwar im Durchschnitt gut funktionieren, aber ebenfalls anfällig für Verzerrungen sind, vergleichbar mit menschlicher Kognition und somit auch ihren Grundannahmen verhaftet sind (vgl. Mittelstadt 2019). Ähnlich wie der Mensch, ten-

dieren die zugrundeliegenden Sprachmodelle dazu die Komplexität der sozialen Realität zu abstrahieren und dabei unter andere, auf quantitative Verlustfunktionen zurückzugreifen, um Entscheidungen zu optimieren (vgl. Carey/Wu 2023; Kamiran/Calders 2009). Diese Verlustmetriken führen häufig zu einer Übervereinfachung, wodurch ein systemischer Unterschied zwischen dem modellierten Bild und der tatsächlichen Realität entsteht – ein Problem der internen Validität. Darüber hinaus leidet die externe Validität vieler KI-Anwendungen. In komplexen, dynamischen soziotechnischen Kontexten können Modelle – die unter bestimmten Bedingungen trainiert wurden – unter realen Einsatzbedingungen versagen, insbesondere wenn diese Bedingungen nicht vollständig abgebildet oder antizipiert wurde (vgl. Grimes/Schulz 2002). Ein anschauliches Beispiel sind Gesichtserkennungssysteme zum Entsperren von Accounts, die bei unzureichend differenziertem Training nicht in der Lage sind, alle Gesichter zuverlässig zu erkennen.

Wird nun unter der vereinfachten Annahme argumentiert, dass digitale Systeme lediglich ein Abbild der analogen Welt darstellen (vgl. Nassehi 2019), ergibt sich daraus die logische Konsequenz, dass auch KI-Entscheidungen nur in dem Maß gerecht oder objektiv sein, wie die zugrundeliegenden analogen, menschengemachten Annahmen und dessen Verarbeitung. Und dennoch: Der Diskurs über KI-Fairness unterscheidet sich von jenem über Entscheidungsgerechtigkeit, wenn Menschen die Akteure sind, was grundlegend mit den unterschiedlichen Ansprüchen an zwei kognitiven Systemen für Entscheidungsfindung und der jeweiligen angenommenen menschlichen Fehlbarkeit in Bezug gebracht werden kann (vgl. Kahneman 2012). Während bei menschlichen Entscheidungen häufig normative und situative Komponenten berücksichtigt werden, folgt die Debatte über algorithmische Fairness anderen Logiken – etwa solchen der mathematischen Gleichbehandlung, der Transparenz oder der Reproduzierbarkeit. Dafür bedarf es einer vorläufigen Reflexion, was Gleichbehandlung ausmacht, welche Parameter für Verständnis transparent gemacht werden müssen oder was es zu wissen gilt wenn Entscheidungen unter gleichen Bedingungen wieder gleich getroffen werden sollen, in Gegenwart alternativer Möglichkeiten. Daraus ergibt sich die Notwendigkeit, KI nicht lediglich als technologische Ergänzung zu Entscheidungen von Menschen zu verstehen, sondern als eigenständiges sozio-technisches Artefakt, das neue ethische und epistemologische Fragen aufwirft und ebenfalls der Entscheidungslogik unterliegt. Dieses grundlegende Dilemma ist es, was uns die Antwort so schwer macht auf das bekannte Trolley-Problem. Würde unsere moralische Bewertung eines Unfalls, bei dem ein autonomes

Fahrzeug drei Senioren anfährt, anstelle eines auf die Straße laufenden Kindes – weil diese Priorisierung als Regel vom Hersteller der KI vorgegeben wurde, gleich ausfallen wie bei einem Menschen am Steuer, der instinktiv gehandelt hat?

In Abbildung 1 ist eine zusammenfassende Darstellung der Charakteristiken von unseren beiden Denksystemen aufgeführt (vgl. ebd.) und wie sich die Qualitäten von KI-Entscheidungen dazu verhalten. KI-Entscheidungen weisen eine hybride Charakteristik auf, die sich nicht eindeutig einem der beiden Systeme zuordnen lässt, sondern Qualitäten beider vereint. Worin genau das Potenzial von KI-Entscheidungen liegt. Aus ökonomischer Perspektive – im Hinblick auf Effizienz- und Funktionalitätsmaximierung – bietet KI klare Vorteile: Sie vereint die Geschwindigkeit und Automatisierung mit der Fähigkeit, komplexe Entscheidungsprobleme in kürzester Zeit bewältigen zu können. Die Frage, inwieweit generative Modelle – mit wachsender Autonomie und Intransparenz ihrer Entscheidungslogik – noch als „bewusst" im Sinne menschlicher Urteilskraft klassifizierbar sind, bleibt offen und bedarf weiterführender Diskussionen, die unter dem Prinzip der Erklärbarkeit bereits allgegenwertig sind.

In diesem Beitrag wird der Argumentation gefolgt, dass das Bewusstsein dem Menschen vorbehalten bleibt und bei KI als prozessuale Quelle lediglich von logischer Intelligenz gesprochen werden kann, die auf Mustererkennung aufbaut, sich lernend weiterentwickeln kann – gemäß der Management Prämisse nach den Superlativen der Effizienz und Funktionalität zu streben – und Probleme nach Regeln löst (vgl. Hildt 2019). So erfordert die algorithmische Grundlage von KI grundsätzlich logische Entscheidungsprozesse syntaktischer Regeln, ohne semantisches Verständnis, erklärt durch die Notwendigkeit expliziter Modellierung und Parametrisierung (vgl. Searle 1980). Darauf aufbauend lässt sich die Frage der Verlässlichkeit bzw. Fehleranfälligkeit bzgl. Fairness als eine Frage der richtigen Steuerung von Parameterwahl, Kausalmodellen und präventive Eingriffsmöglichkeiten im Designprozess im Umgang mit Unschärfe und Ambiguität umformulieren.

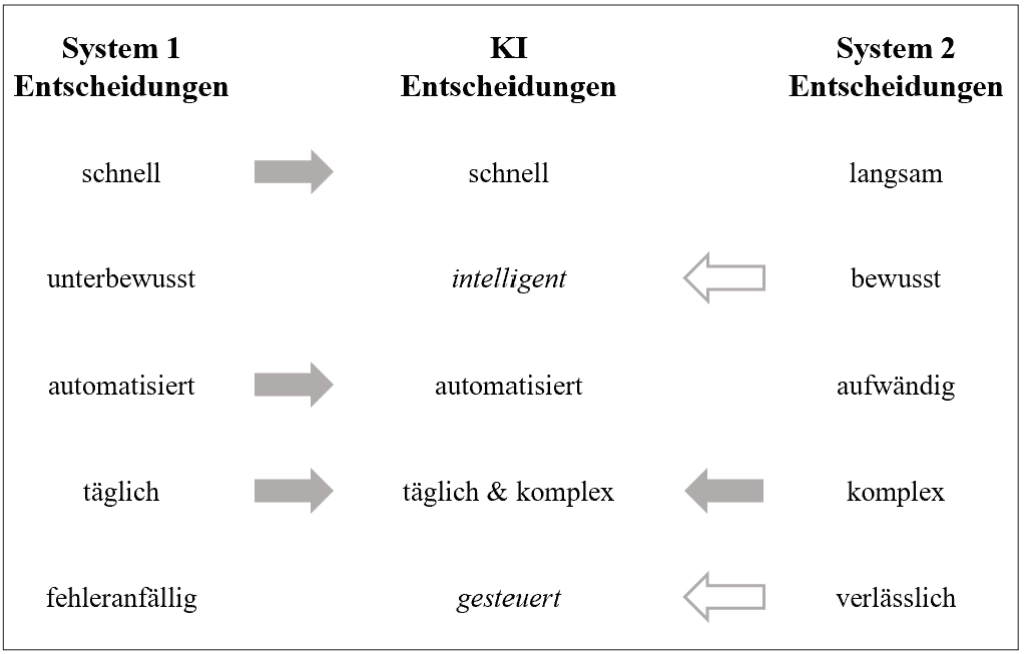

ABBILDUNG 1: ENTSCHEIDUNGSLOGIK MENSCH UND KI
(QUELLE: EIGENE DARSTELLUNG IN ANLEHUNG AN KAHNEMAN 2012)

3. Reflexionsleitfaden KI-Fairness in die Unternehmenspraxis bringen

Der überwiegende Anwendungsfall für KI heute ist die Übernahme von bestehenden Aufgaben (z. B.: Bewerbungsprozesse, Kreditvergaben, Kundenservice oder Routenplanung) deren Parameter bereits gesetzt sind (z. B.: Preise, Fristen, Qualitäten). Systeme in Unternehmen, die über Jahre gewachsen sind und was sich bewährt hat, wird erhalten, denn auch Unternehmen sind autopoietische Systeme. Wenn keine Notwendigkeit besteht, wie hohe Fluktuation, braucht es einen externen Faktor – wahrscheinlich sind die meisten Fluktuationswellen ebenfalls durch externe Wirkung bedingt – der Veränderung ins System bringt (vgl. Luhmann 2000). Abstrahiert kann diese externe Wirkung auf der Ebene der Kriterien, Kanäle oder Köpfe stattfinden und meistens braucht es eine Kombination aus allen drein. Darauf aufbauend wird im nachfolgend

der Kanal Datenfluss als Inputquelle und Outputgröße für KI betrachte und ein Reflexionsprozess für Kriterien, orientiert an dem Datenlebenszyklus von Jones (2019), ausgeführt. Daraus folgt abschließend die Implikation für den Wandel auf der Ebene der Köpfe.

Der Datenlebenszyklus setzt an der Quelle an – dem Erhebungsprozess von Daten die für KI verwendet werden. Hierbei wird schrittweise analysiert, welche strukturellen Entscheidungen bereits in der Datenerhebung getroffen werden und wie diese Fairness beeinflussen. Schrittweise wird präzisiert, welche Daten einbezogen werden bis zur Nutzung und Interpretation der erhobenen Daten. Auch hier ist eine systemische Reflexion erforderlich, um Verzerrungen zu vermeiden. Ob es nun um den Anspruch an Fairness in der Entstehung oder Nutzung der Daten geht – „How data comes to be" und „How data comes to be used" (Jones 2019: 12) – im Fokus der nachfolgenden Betrachtung soll ein Reflexionsprozess des Bewusstmachens entstehen. Die technische Ebene der Umsetzbarkeit von Fairness – bezüglich dessen bereits langjährige und weitreichende wissenschaftliche Erkenntnisse bestehen (z. B. Friedman/Nissenbaum 1996) – wird an dieser Stelle nur indirekt betrachtet. Die nachfolgenden Schritte sind auf der Ebene der digital verantwortlichen Unternehmensführung zu betrachten – „ganzheitliches verantwortungsbewusstes Handeln von Unternehmen in Bezug auf die Auswirkungen digitaler Technologien und Daten" (Carl et al. 2025: 10), mit dem Ziel steuernder Elemente zur Erkennung, Abschwächung und Kontrolle von Verzerrungen zu bieten.

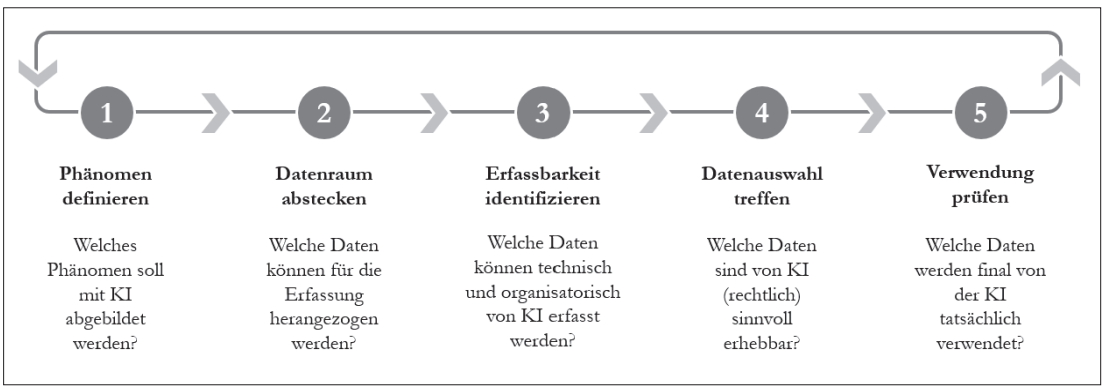

ABBILDUNG 2: REFLEXIONSLEITFADEN
(QUELLE: EIGENE DARSTELLUNG IN ANLEHUNG AN JONES 2019)

Phänomen definieren

Zunächst gilt es zu bestimmen, „welches Phänomen mit KI abgebildet werden soll". Dafür muss konzeptionell definiert sein, welches Phänomen zu welchem Zweck von KI betrachtet bzw. nachgebildet wird. Nur wenn der Zweck der Datenerhebung präzise definiert ist, lassen sich zweckgebundene und damit zielführende Daten erheben. Mit Blick auf Fairness sollte Klarheit darüber hergestellt werden, ob das Phänomen, wie es in der Vergangenheit abgebildet wurde auch in der Zukunft abgebildet werden soll oder es Anpassungsbedarf an weitere und zukünftige Bedingungen gemäß dem Anspruch an Fairness gibt. Die Antwort auf diese Frage ist eine entscheidende Weichenstellung für den weiteren Datenanalyseprozess. Unter der Annahme, dass das Digitale nur eine Abbildung vom Analogen ist, muss nachfolgend interveniert werden, wenn Ergebnisse aus der Vergangenheit nicht in die Zukunft skaliert werden sollen (vgl. Nassehi 2019), die Steuerung der Daten also künstlich angepasst werden muss oder die Möglichkeit dafür eingeräumt werden sollte. Herausfordernd unter der Allgegenwertigkeit von KI und der nachträglichen Formbarkeit des Anwendungszwecks (vgl. Lobschat et al. 2021).

- Fallbeispiel 1: Im Rahmen eines unternehmensweiten Talentmanagement-Programms soll eine KI eingesetzt werden, um Mitarbeitende mit besonders hohem Potenzial für zukünftige Führungsaufgaben zu identifizieren. Ziel ist es, datenbasiert und vorausschauend Talente zu erkennen, um diese gezielt und frühzeitig zu fördern. Anders als zuvor sollen Mitarbeitende aus allen Standorten und unabhängig vom Bildungsweg für Führungspositionen evaluiert werden.
- Fallbeispiel 2: Für die Einführung einer neuen Produktlinie plant das Unternehmen den Einsatz von KI für die Lieferantenauswahl. Die KI soll ein belastbares, vielfältiges und nachhaltiges Portfolio an Lieferanten zusammenstellen, das sowohl ökonomischen Zielwerten als auch zukünftig ökologischen und sozialen Anforderungen gemäß neuer Lieferkettengesetzte gerecht wird.

Datenraum abstecken

Anschließend stellt sich die Frage, „welche potenziellen Daten können für die Erfassung und Beschreibung des Phänomens herangezogen werden". Hierbei wird ein möglichst breites Spektrum potenzieller relevanter Daten identifiziert, ohne diese unmittelbar hinsichtlich ihrer Relevanz oder Erfassbarkeit zu bewerten. An dieser Stelle den Blick möglichst weit zu öffnen ist notwendig für Diversifizierung und Ganzheitlichkeit im entstehenden Modell. Kulturelle Eigenschaften, limitierende Rahmenbedingungen oder persönliche Spezifika produzieren Erfahrungswerte, die zu

beschränkter Reproduktion von Beschreibungen führen. Um diese Autopoiesis zu unterbrechen, bedarf es meist einer Intervention von außen, ein Reevaluieren der zugrundeliegenden Parameter die Entscheidungsabläufe ausmachen (vgl. Luhmann 2000). Es gilt zu reflektieren, wer mit welcher Perspektive Daten definiert, wo die Grenzen liegen können und was ergänzend hinzugenommen werden soll. Ein kritischer Punkt ist, dass technische Systeme menschliche Entscheidungsprozesse nur bedingt abbilden können (vgl. Friedman/Nissenbau 1996). So spielen in Managemententscheidungen beispielweise häufig Bauchgefühle eine Rolle, die nicht durch explizite Daten erfassbar sind. Solche Limitationen müssen im Erfassungsprozess bewusst definiert und berücksichtigt werden (vgl. Shahbazi et al. 2023).

- Fallbeispiel 1: Zur Einschätzung des Führungspotenzials könnten die folgenden quantifizierbaren und qualitativen Datenpunkte herangezogen werden: Ergebnisse vergangener Leistungsbeurteilungen, Teilnahme und Abschluss von Weiterbildungsmaßnahmen, Dauer der Unternehmenszugehörigkeit, Anzahl und Qualität erfolgreich abgeschlossener Projekte sowie kollegiales und vorgesetztenbasiertes Feedback. Da sich die Unternehmensführung für Krisenzeiten resilient aufstellen will, wird von der Strategieabteilung zusätzlich gefordert, dass die Ambiguitätstoleranz der Mitarbeitenden miterfasst wird. Im Zug der digitalen Transformation des Kerngeschäfts soll außerdem die Affinität zu und der Umgang mit digital Technologien berücksichtigt werden.
- Fallbeispiel 2: Für die Beurteilung potenzieller Lieferanten könnten unter anderem folgende Datenpunkte als Grundlage dienen: historische Lieferzuverlässigkeit (Pünktlichkeit und Flexibilität), Qualität der gelieferten Wahre, Preisentwicklung und Kostenstabilität, Standort und logistische Anbindung, Einhaltung von Nachhaltigkeits- und Sozialstandards sowie Diversitätskriterien. Seitens der Risikoabteilung wird ergänzt, dass potenzielle Auswirkungen geopolitischer Entwicklungen – wie Kriege oder veränderte Zollpolitik – auf Lieferwege mit zu berücksichtigen sind.

Erfassbarkeit identifizieren

Innerhalb des definierten Möglichkeitsraum muss differenziert werden, „welche Daten technisch und organisatorisch von KI erfasst werden können". Mit der Eingrenzung sollte reflektiert werden, ob Daten potenzielle „Pre-Existing Biases" aus vergangenen Strukturen enthalten, die zu Stichprobenverzerrungen führen (vgl. Barocas et al. 2023). Eventuell wurden bestimmte Gruppen

historisch unzureichend erfasst oder waren unterrepräsentiert, was zum damaligen Zeitpunkt Realitäten stimmig abgebildet hat, skaliert ins Heute jedoch zu falschen Schlussfolgerungen führen kann. Eine kritische Reflexion der Datenquellen und ihrer Verteilungen ist notwendig, um ein möglichst unverzerrtes Bilder der aktuellen Realität zu gewährleisten (vgl. Mehrabi et al. 2021; Olteanu et al. 2019; Veale/Binns 2017). In einem iterativen Abgleich mit den vorläufigen Schritten muss erfasst werden, welche Daten, ausgehend vom zu beschreibenden Phänomen, notwendig sind und welche Daten eine hinreichende Beschreibung ermöglichen.

- Fallbeispiel 1: Die kritische Überprüfung der identifizierten Daten zeigt, dass die verwendeten Kriterien möglicherweise nicht alle relevanten Aspekte abdecken. Soziale Kompetenzen lass sich schwer quantifizieren und können daher unterbewertet werden. Bei der Erfassbarkeit der ausgewählten Daten zeigt sich, dass kollegiales Feedback nicht standortübergreifend erhoben wird und das eingesetzte Mitarbeitendentools nicht für jeden Anwendungsfall eine realitätsgetreue Abbildung ermöglichen. Zusätzlich findet informelle Rückmeldung, die im Arbeitsalltag eine große Rolle spielt, oftmals keinen Eingang in strukturierte Erfassungssysteme.
- Fallbeispiel 2: In der Lieferantenbewertung zeigt sich, dass Unterschiede im Digitalisierungsgrad zu Verzerrungen in der Bewertung führen. Besonders kleinere Lieferanten und Partner aus dem globalen Süden weisen vermehrt eine geringere Datenverfügbarkeit und -qualität auf. In der Bewertung muss sichergestellt werden, dass ein Nichtvorhandensein von Daten nicht mit einem Nichterfüllen gleichgesetzt wird und Annahmen, die zur Schließung von Datenlücken verwendet werden, transparent gemacht werden.

Datenauswahl treffen

Anschließend erfolgt eine Bewertung, „welche Daten von KI (rechtlich) sinnvoll erhebbar sind", wobei Aspekte der Datensparsamkeit, Repräsentativität und Verzerrungsfreiheit eine zentrale Rolle spielen, u.a. reguliert durch Datenschutzgrundverordnungen (DSGVO in der EU). Dabei müssen insbesondere „Self-Selection Biases" und „Selection Biases" berücksichtigt werden. Ersteres tritt auf, wenn Daten allein aufgrund ihrer Verfügbarkeit genutzt werden, ohne kritisch zu hinterfragen, ob sie für das betrachtete Phänomen repräsentative sind. Letzteres betrifft Verzerrungen, die durch die Instanz entstehen, welche die Daten sammelt. Resultierende problematische Kausalitätsannahmen müssen vermieden werden, da fehlerhafte Schlussfolgerungen zu unfairen oder diskriminierenden Entscheidungen führen können. Ein klassisches Beispiel für einen Self-

Selection Bias, der zu ungewollten Kausalitäten führte, ist das Recruitingtool von Amazon. Da bisher überwiegend Männer für die ausgeschriebene Funktion eingestellt waren schlussfolgerte die KI, dass es ein relevantes Kriterium ist und sortierte Frauen aus. Auch Verzerrungen durch den Einsatz von digitalen Technologien und die digitale Verarbeitung von Daten müssen reflektiert werden, da die Anwendung eines Modells oder die Kategorisierung sensibler Daten zu unerwarteten Verzerrungen führen kann (vgl. Olteanu et al. 2019; Suresh/Guttag 2019).

- Fallbeispiel 1: Nicht alle ausgewählten Daten lass sich für die Kandidatenauswahl ohne weiteres erfassen. Soziale Kompetenzen, die für eine Führungsposition grundlegend sind, sind schwer quantifizierbar und können daher unterbewertet werden. Die Nutzung von Überwachungstechnologien am Arbeitsplatz, zur Leistungserfassung wird zunehmend gebräuchlicher, stellt jedoch gravierende Eingriffe in die Privatsphäre dar – teilweise rechtlich eingeschränkt und generell ethisch kritisch zu betrachten. Darüber hinaus sind sensible Daten wie Elternzeit oder Krankheitsausfälle aus ethischen und teils datenschutzrechtlichen Gründen nur eingeschränkt verwendbar und beeinflussen die Kennzahl der kumulierten Unternehmenszugehörigkeit.
- Fallbeispiel 2: Bei der Auslegung von Auswahlkriterien kann es zu systematischen Benachteiligungen – etwa auf Grund regionaler Unterschiede im Verständnis von ökologischen und sozialen Standards, die bei dichotomer Erfassung ein zu undifferenziertes Bild von Erfüllbarkeit erzeugen. Zudem können bestimme Kommunikationsdaten aus Datenschutzgründen, abhängig vom Herkunftsland, nicht von allen Lieferanten gleichermaßen erhoben oder verarbeitet werden, was zum Ausschluss von einzelnen Partnern führte.

Verwendung prüfen

Im letzten Schritt wird der Präzisierung nachgegangen, „welche Daten werden final von der KI tatsächlich verwendet“. Die Zuverlässigkeit der Datenerhebung ist hierbei von entscheidender Bedeutung, da Fehler oder Ausfälle in der Soft- und Hardware zu Abweichungen zwischen erfassten und registrierten Daten führen können. Zudem können frühere Entscheidungen im Erfassungsprozess unbeabsichtigte Auswirkungen auf das finale Ergebnis haben, sodass die angestrebte objektive Darstellung des Phänomens nicht gewährleistet ist. Dieser finale Schritt ist eng mit dem Anspruch an Erklärbarkeit (Explainable AI [XAI]) verbunden, welcher jedoch nicht als Kausalannahme für Fairness gesetzt werden kann. Sprich KI-basierter Entscheidungsprozess, gesteuert gemäß dem Fairnessprinzip bedarf einer weiteren Reflexion der Fairnessqualität, die am

Ende erzielt wird. In den meisten Fällen liegt der Fokus stark auf Fairness im Sinne von „Equality of Opportunity", alle Optionen bekommen dieselbe Chance ausgewählt zu werden durch repräsentative Daten zu Beginn (vgl. Kodelja 2016; Roemer 1998).

Was dahinter jedoch noch Berücksichtigung erfahren sollte – besonders wenn wir neben den klassischen Unternehmensparametern, Effizienz und Funktionalität, weitere Ansprüchen Berücksichtigung einräumen wollen – ist KI-Fairness gemäß „Equality of Effort" (vgl. Huang et al. 2020). Hierbei wird auf die Inputgröße geschaut, die notwendig ist, um ein gewisses Ergebnis zu erreichen, wobei der Superlativ davon schnell in ein meritokratisches System abbiegen kann. Mit Blick auf das faire Ergebnis selbst – hier ist der größte Anspruch der idealistisch in den Raum gestellt wird gemäß einer Solidaritätsgesellschaft, soziale Gerechtigkeit, die zunehmend zum Verantwortungsportfolio von Unternehmen gezählt wird – dabei geht es um Fairness als „Equality of Outcome" (vgl. Kodelja 2016). Ein Anspruch der schnell als philanthropische Verantwortung abgetan werden kann, aber durchaus die zunehmende Stakeholderlandschaft ins Gespräch ruft, die in einer Gleichzeitigkeit Antworten auf die Frage „Fair für wen?" abverlangt. Daten nicht nur als Ein- und Ausschluss betrachten, sondern als abhängige Größe, die in ihrer Gesamtheit betrachtet werden soll, um eine gut informierte Entscheidung zu erzielen.

- Fallbeispiel 1: Da das Unternehmen eine KI-Lösung für das Talentmanagement eingekauft hat, ist die eigene Nachvollziehbarkeit des Fairnessprinzips begrenzt und Entscheidungsträger*innen sind von den Bedingungen des Anbieters abhängig. Um zu identifizieren, ob die getroffene Datenauswahl im Sinne einer fairen, nachvollziehbaren und erklärbaren Entscheidung für die Förderung bestimmter Talente steht, werden jährliche Ergebniskontrollen etabliert. Da jedoch unerwünschte Muster meist erst nach ausreichender Datenmenge erkennbar sind, wird der Talentmanagementprozess als enger „Human-on-the-Loop" aufgebaut, um die KI-Entscheidungen unterjährig mit den definierten Parametern zur Beschreibung des Phänomens iterativ abzugleichen.
- Fallbeispiel 2: Das Unternehmen hat für die Lieferantenauswahl ihre eigene KI entwickelt und über multiple Ebenen in den Prozess festgelegt, wie die Kriterien – festgelegt in den Schritten zuvor – bei der Auswahl Anwendung finden sollen. Um jedoch sicherzugehen, dass die Annahmen, übersetzt in Kriterien das Wertesystem des Unternehmens abbilden, wird dennoch eine regelmäßige stichprobenbasierte Ergebnisprüfung durchgeführt, iterativ mit Expert*innen der technischen Weiterentwicklung und dem Feedback der Marketingabteilung über Konsument*innenverhalten.

4. *Wirkungsgröße KI schafft neue Relevanz von Fairness*

Das vorgestellt fünfstufige Modell dient als strukturierte Orientierung für Entscheidungstragende in Unternehmen, um Fairness ganzheitlich in den gesamten Datenprozess hinter KI zu verankern – von der Erhebung über die Verarbeitung bis hin zur Nutzung. Ziel ist es, durch bewusste und reflektierte Entscheidungen in jeder Phase des Prozesses Verzerrungen zu erkennen, zu minimieren und so eine gerechtere Datenverarbeitung zu ermöglichen. Nur so lassen sich (unbeabsichtigte) negative Auswirkungen von unfairer KI auf Individuen und Gesellschaft reduzieren und vermeiden (vgl. Shelby et al. 2023). Es gilt, kognitive Automatismen, die auf erlernten Heurisiken und Vorannahmen basieren, nicht unreflektiert von KI übernehmen zu lassen. Vielmehr sollte das reflektierte System der menschlichen Entscheidungsfindung als Grundlage dienen, um bestehende Entscheidungsprozesse kritisch zu hinterfragen und gegebenenfalls neu auszurichten. Besonders herausfordernd gestaltet sich dies, wenn nicht eindeutig priorisiert werden kann, wem gegenüber Fairness zu gewährleistet ist – und wenn gegebenenfalls Kriterien potenziell im Konflikt zueinanderstehen (vgl. Sanderson et al. 2023; van Nood/Yeomans 2021). Im Kontext des Fallbeispiel 1 zeigt sich dies exemplarisch: Die Einführung von Frauenquoten führt zu Zielkonflikten, die sich nicht immer vollständig auflösen lassen. Die Erhöhung des Frauenanteil in Führungspositionen und gleichzeitig die objektiv beste Auswahl zu treffen stellt, sich nicht immer als kompromissfreie Entscheidung dar und ist gegebenenfalls unfair anderen Kandidaten gegenüber.

Der mit diesem Beitrag intendierte Impuls richtet sich an die unternehmerische Steuerungsebene, um dort ein Bewusstsein zu schulen für das Analoge als Grundlage für das Digitale und die damit einhergehende Wirkungsform auf „faire Daten rein – faire Entscheidungen raus". Dieses Bewusstsein darf jedoch nicht zu einer trügerischen Sicherheit führen. Denn auch wenn fairen analoge Grundlagen eine wichtige Voraussetzung darstellen, hebt das Digitale die Wirkungsgrößen auf ein Neues, teils noch nicht absehbares Ausmaß. KI entwickelt sich mit exponentieller Geschwindigkeit und eskaliert damit die Bedeutung der neuen Währung, Daten. Die ursprüngliche Zweckgebundenheit von KI-Entwicklung ist zunehmend obsolet, bei gleichzeitigen Anwendungsbereichen die allgegenwärtig sind, zunehmend integriert in nahezu alle Unternehmensebenen, Branchen und Lebensbereiche (vgl. Lobschat et al. 2021). Die Reichweite macht nicht an System- oder Ländergrenzen halt und fordert enge Zusammenarbeit zwischen unterschiedlichen Wertesystemen, politischen Strukturen und Unternehmen.

Alvarez, J. M. / Colmenarejo, A. B. / Elobaid, A. / Fabbrizzi, S. / Fahimi, M. / Ferrara, A. / Ghodsi, S. / Mougan, C. / Papageorgiou, I. / Reyero, P. (2024): Policy Advice and Best Practices on Bias and Fairness in AI, in: Ethics and Information Technology, Jg. 26 / Nr. 2, 1–26, DOI: 10.1007/s10676-024-09746-w.

Beutel, A. / Chen, J. / Doshi, T. / Qian, H. / Woodruff, A. / Luu, C. / Kreitmann, P. / Bischof, J. / Chi, E. H. (2019): Putting Fairness Principles into Practice: Challenges, Metrics, and Improvements. 19: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, Honolulu, 453–459, DOI: 10.1145/3306618.331423.

Barocas, S. / Hardt, M. / Narayanan, A. (2023): Fairness and Machine Learning. Limitations and Opportunities, Cambridge: MIT Press.

Carey, A. N. / Wu, X. (2023): The Statistical Fairness Field Guide: Perspectives from Social and Formal Sciences, in: AI and Ethics, Jg. 3 / Nr. 1, DOI: 10.1007/s43681-022-00183-3.

Carl, K. Valerie / Blachetta, M. / Bilfinger, B. / Brink, A. / Cerulli-Harms, A. K. / Christ, A.-S. / Dietl, K. / Dörr, S. / Hauer, M. P. / Hinz, M. /Jänig, J.-R. / Keilhofer, K. / Kurtz, C. / Mihale-Wilson, C. / Mueller, B. / Przybylek, B. A. / Schaufler, N. / Schlegel, I. K. / Scholz, M. / von Roesgen, F. / Wößner, J. (2025): Establishing Common Ground in Corporate Digital Responsibility, in: Proceedings Informatik Festival 2025 [im Erscheinen].

Christ, A.-S. (2025): Corporate Governance for Digital Responsibility. A Company Study, Münster: Proceedings of the 20th International Conference on Wirtschaftsinformatik.

Christ, A.-S. / Mueller, B. / Lobschat, L. (2024): Corporate Digital Responsibility: Achieving Synthesized Coherence with CSR Learnings on Corporate Governance, Bangkok: Proceedings of the 45 International Conference on Information Systems.

Crootof, R. / Kaminski, M. E. / Price, W. / Nicholson, I. I. (2023): Humans in the Loop, in: Vanderbilt Law Review, Vol. 76 / No. 6, 429–510, DOI: 10.2139/ssrn.4066781.

CDR-Initiative (2025): Unsere Ziele – Der CDR-Kodex. Die 9 Prinzipien des CDR Kodex, URL: https://cdr-initiative.de/kodex (aufgerufen am: 20/05/2025).

Friedman, B. / Nissenbaum, H. (1996): Bias in Computer Systems, in: ACM Transactions on Information Systems (TOIS), Vol. 14 / No. 3, 330–347, DOI: 10.1145/230538.23056.

Grimes, D. A. / Schulz, K. F. (2002): Bias and Causal Associations in Observational Research, in: The Lancet, Vol. 359 / No. 9, 302, 248–252, DOI: 10.1016/S0140-6736(02)07451-2.

Hildt, E. (2019): Artificial Intelligence: Does Consciousness Matter?, in: Frontiers in Psychology, Jg. 10, Artikel 1535, DOI: 10.3389/fpsyg.2019.01535.

Huang, W. / Wu, Y. / Zhang, L. / Wu, X. (2020): Fairness Through Equality of Effort., Taipei: Companion Proceedings of the Web Conference 2020, 743–751, DOI: 10.1145/3366424.338355.

John-Mathews, J.-M. / Cardon, D. / Balagué, C. (2022): From Reality to World. A Critical Perspective on AI Fairness, in: Journal of Business Ethics, Vol. 178 / No. 4, 945–959, DOI: 10.1007/s10551-022-05055-8.

Jones, M. (2019): What we Talk About When We Talk About (big) Data, in: The Journal of Strategic Information Systems, Vol. 28 / No. 1, 3–16, DOI: 10.1016/j.jsis.2018.10.005.

Kahneman, D. (2012): Schnelles Denken, Langsames Denken, München: Siedler Verlag.

Kamiran, F. / Calders, T. (2009): Classifying Without Discriminating, Karachi: 2nd International Conference on Computer, Control and Communication, IEEE, DOI: 10.1109/IC4.2009.4909197.

Kattnig, M. / Angerschmid, A. / Reichel, T. / Kern, R. (2024): Assessing Trustworthy AI: Technical and Legal Perspectives of Fairness in AI, in: Computer Law & Security Review, Jg. 55, Artikel 106053, DOI: 10.1016/j.clsr.2024.106053.

Kodelja, Z. (2016): Equality of Opportunity and Equality of Outcome, in: Center for Educational Policy Studies Journal, Jg. 6 / Nr. 2, 9–24, DOI: 10.26529/cepsj.85.

Krakowski, S. / Luger, J. / Raisch, S. (2023): Artificial Intelligence and the Changing Sources of Competitive Advantage, in: Strategic Management Journal, Jg. 44 / Nr. 6, 1425–1452, DOI: 10.1002/smj.3387.

Lobschat, L. / Mueller, B. / Eggers, F. / Brandimarte, L. / Diefenbach, S. / Kroschke, M. / Wirtz, J. (2021): Corporate Digital Responsibility, in: Journal of Business Research, Jg. 122, 875–888, DOI: 10.1016/j.jbusres.2019.10.006.

Luhmann, N. (2000): Organisation und Entscheidung, Wiesbaden: Springer Fachmedien.

Madaio, M. / Egede, L. / Subramonyam, H. / Vaughan, J. / Wallach, H. (2022): Assessing the Fairness of AI Systems: AI Practitioners' Processes, Challenges, and Needs for Support, New York: Proceedings of the ACM on Human-Computer Interaction, Jg. 6 / Nr. CSCW1, 1–26, DOI:10.1145/3512899.

Mehrabi, N. / Morstatter, F. / Saxena, N. / Lerman, K. / Galstyan, A. (2021): A Survey on Bias and Fairness in Machine Learning, in: ACM Computing Surveys (CSUR), Jg. 54 / Nr. 6, 1–35, DOI: 10.1145/3457607.

Mittelstadt, B. (2019): Principles Alone Cannot Guarantee Ethical AI, in: Nature Machine Intelligence, Jg. 1 / Nr. 11, 501–507, DOI: 10.1038/s42256-019-0114-4.

Mueller, B. (2022): Corporate Digital Responsibility, in: Business and Information Systems Engineering, Jg. 64 / Nr. 5, 689–700, DOI: 10.1007/s12599-022-00760-0.

Nassehi, A. (2019): Muster: Theorie der Digitalen Gesellschaft, München: C.H. Beck.

Olteanu, A. / Castillo, C. / Diaz, F. / Kıcıman, E. 2019 : Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries, in: Frontiers in Big Data, Jg. 2 / Nr. 13, DOI: 10.3389/fdata.2019.00013.

Pant, A. / Hoda, R. / Tantithamthavorn, C. / Turhan, B. (2024): Navigating Fairness: Practitioners' Understanding, Challenges, and Strategies, in: Empirical Software Engineering, Jg. 30 / Nr. 4, 102, DOI: 10.1007/s10664-025-10650-0.

Roemer, J. E. (1998): Equality of Opportunity, London: Harvard University Press.

Sanderson, C. / Douglas, D. / Lu, Q. / Schleiger, E. / Whittle, J. / Lacey, J. / Newnham, G. / Hajkowicz, S. / Robinson, C. / Hansen, D. (2023): AI Ethics Principles in Practice: Perspectives of Designers and Developers, in: IEEE Transactions on Technology and Society, Jg. 4 / Nr. 2, 171–187, DOI: 10.1109/TTS.2023.3257303.

Searle, J. R. (1980): Minds, Brains, and Programs, in: Behavioral and Brain Sciences, Jg. 3 / Nr. 3, 417–424, DOI: 10.1017/S0140525X00005756.

Shahbazi, N. / Lin, Y. / Asudeh, A. / Jagadish, H. (2023): Representation Bias in Data: A Survey on Identification and Resolution Techniques, in: ACM Computing Surveys, Jg. 55 / Nr. 13.

Shelby, R. / Rismani, S. / Henne, K. / Moon, A. / Rostamzadeh, N. / Nicholas, P. / Yilla-Akbari, N. M. / Gallegos, J. / Smart, A. / Garcia, E. (2023): Sociotechnical Harms of Algorithmic Systems: Scoping a Taxonomy for Harm Reduction, Montreal, in: Rossi, F. / Das, S. / Davis, J. / Kay, F.-B. / John, A. (Hrsg.): AIES'23: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, New York: Association for Computing Machinery, 723–741, DOI: 10.1145/3600211.3604673.

Schiff, D. / Rakova, B. / Ayesh, A. / Fanti, A. / Lennon, M. (2021): Explaining the Principles to Practices Gap in AI, in: IEEE Technology and Society Magazine, Vol. 40 / No. 2, 81–94, DOI: 10.1109/MTS.2021.3056286.

Suresh, H. / Guttag, J. (2019): A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle, New York: Proceedings of the 1st ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization, New York: Association for Computing Machinery, DOI: 10.1145/3465416.348330.

van Nood, R. / Yeomans, C. (2021): Fairness as Equal Concession: Critical Remarks on Fair AI, in: Science and Engineering Ethics, Jg. 27 / Nr. 73 DOI: 10.1007/s11948-021-00348-z.

Veale, M. / Binns, R. (2017): Fairer Machine Learning in the Real World: Mitigating Discrimination without Collecting Sensitive Data, in: Big Data & Society, Jg. 4 / Nr. 2, DOI: 10.1177/2053951717743530.

Zhang, W. (2024): AI Fairness in Practice: Paradigm, Challenges, and Prospects, in: AI Magazine, Jg. 45 / Nr. 3, 386–395, DOI: 10.1002/aaai.12189.