Fairness bei KI erfordert Führung[1]

Warum Führung heute entscheidet, wie gerecht die Welt morgen ist

Saskia Dörr

*1. Die neue Macht der Empfehlungssysteme – Warum Fairness Chef*innensache ist*

Empfehlungssysteme, Credit-Scoring-Algorithmen, Chatbots, Recruiting-Tools oder Segmentierungstechnologien beeinflussen heute alltägliche Lebensentscheidungen – oft automatisiert, vielfach unsichtbar, aber mit tiefgreifenden Konsequenzen. Sie entscheiden mit darüber, wer Zugang erhält: zu Informationen, Jobs, Krediten, Bildung oder sozialen Netzwerken. Algorithmische Systeme gestalten so nicht nur Prozesse, sondern auch Machtverhältnisse: Sie bestimmen mit, wer gesehen, gehört und berücksichtigt wird – im Hier und Jetzt, aber auch mit langfristigen Folgen. Denn jede algorithmische Entscheidung kann bestehende Ungleichheiten verfestigen, verschärfen oder abbauen. Fairness wird damit zur strategischen Führungsfrage. Sie ist nicht nur ein verfassungsrechtliches Ideal, sondern ein zentrales Prinzip nachhaltiger Unternehmensverantwortung. In einer Zeit, in der Künstliche-Intelligenz (KI)-Systeme über Chancen und Teilhabe mitentscheiden, braucht es vorausschauendes Handeln – als Teil von Verantwortung, Resilienz und Zukunftsfähigkeit. Dass algorithmische Systeme diskriminieren können, ist längst belegt, z. B:

- Eine KI im Recruiting, die auf verzerrten Trainingsdaten basiert, sortiert systematisch bestimmte Bewerbergruppen aus (vgl. Dastin 2018).
- Eine Empfehlungs-KI zeigt Anzeigen nur bestimmten demografischen Gruppen – und verstärkt so ungleiche Sichtbarkeit (vgl. Lambrecht/Tucker 2019).

---

1 Teile dieses Fachtexts und eine Grafik wurden unter Verwendung generativer KI-Tools erstellt (ChatGPT 4.0 für Recherche und Textentwurf, Midjourney für Grafikdesign). Alle Ergebnisse wurden von mir fachlich überprüft und bearbeitet.

- Eine Sprach-KI bevorzugt bestimmte Sprechweisen – und diskriminiert damit implizit all jene, die von kulturellen Normen abweichen (vgl. Bajorek 2019).

Die perfide Qualität algorithmischer Diskriminierung liegt in ihrer Kodierung und Skalierbarkeit: Was einmal verzerrt ist, kann millionenfach automatisiert reproduziert werden. Fairness in KI ist daher keine technische Detailfrage, sondern ein normatives, wirtschaftliches und gesellschaftliches Schlüsselthema. Gerade deshalb sollten sich Unternehmen bewusst machen, welche konkreten Folgen mangelnde Fairness in KI-Systemen nach sich ziehen kann – nicht nur für Betroffene, sondern auch für die Organisation selbst:

- Reputationsrisiken: Diskriminierende Systeme führen zu Skandalen – in Medien, Öffentlichkeit und Zivilgesellschaft.
- Rechtsrisiken: Neue Gesetze wie der EU-AI Act verlangen diskriminierungsfreie, nachvollziehbare KI-Systeme im Hochrisikobereich – Verstöße ziehen Bußgelder und Klagen nach sich.
- Vertrauensverlust: Nutzer*innen, Kund*innen und Mitarbeitende erwarten nicht nur funktionierende, sondern faire Technologien. Wer dieses Vertrauen verliert, gefährdet Kundenbindung, Markenimage und Innovationskraft.

Gleichzeitig eröffnet sich darin ein strategisches Handlungsfeld: Unternehmen, die Fairness aktiv gestalten, positionieren sich als Vorreiter verantwortungsvoller KI. Sie sichern nicht nur regulatorische Konformität, sondern gewinnen Vertrauen, Differenzierung und Zukunftsfähigkeit. Die zentrale Managementfrage lautet: Welchen Stellenwert messen wir Fairness bei – angesichts von Effizienzdruck, technischer Komplexität und unterschiedlichen Vorstellungen von Gerechtigkeit? Führung bestimmt maßgeblich die Balance zwischen technischer Machbarkeit, wirtschaftlicher Effizienz und sozialer Verantwortung (vgl. Abbildung 1). Fairness entsteht nicht automatisch durch den Einsatz von KI – sie muss aktiv und bewusst gestaltet werden.

ABBILDUNG 1: „WELCHES GEWICHT GEBEN WIR FAIRNESS?“
(QUELLE: EIGENE DARSTELLUNG KI-GENERIERT MIT HILFE VON MIDJOURNEY)

Dieser Beitrag stellt einen Ausgangspunkt einer handlungsorientierten Auseinandersetzung dar. Es sensibilisiert für die normative Relevanz von Fairness in KI und skizziert die zentralen Managementherausforderungen:

- Wie können technische, organisatorische und kulturelle Hebel kombiniert werden, um Fairness praktisch zu verwirklichen?
- Welche Dilemmata, Zielkonflikte und unbequemen Wahrheiten müssen dabei ausgehalten werden?
- Welche Rolle spielt Corporate Digital Responsibility (CDR) als Kompass und Struktur für faire KI-Transformation?

Dabei soll der Text dazu dienen auf Basis des Verständnisses der Komplexität Mut zu machen – und machen Mut zur Positionierung, zur Reflexion und zur Gestaltung eines auf Fairness ausgerichteten Management- und Führungssystems. Denn der Umgang mit KI entscheidet, wie fair die Welt heute ist – und ob sie morgen gerechter wird.

2. Was bedeutet „fair" – und warum ist es so schwer, es algorithmisch umzusetzen?

„Das ist unfair!" – diesen Satz haben die meisten von uns bereits in der Kindheit häufig verwendet. Das Wort klingt vertraut, doch kennen wir das komplexe, vielschichtige und kulturell geprägte Konzept von Fairness wirklich? In der Kindheit mag Fairness mit gleich verteilten Bonbons assoziiert sein. Im Unternehmenskontext hingegen gilt sie häufig als Aufgabe der Personalabteilung und wird mit Themen wie Antidiskriminierung, Gleichstellung oder der Förderung von Diversität unter Beschäftigten in Verbindung gebracht. In der algorithmischen Welt hingegen ist Fairness weder selbstverständlich noch eindeutig definiert. Sie ist alles andere als „kodierbar" und kein Automatismus, sondern ein vielschichtiges Gestaltungsziel. Ein fundiertes Verständnis dafür ist zentral – insbesondere im Spannungsfeld zwischen normativen Erwartungen und technischen Umsetzungen. Der folgende Abschnitt stellt daher ausgewählte Fairness-Konzepte vor – einschließlich ihrer philosophischen Grundlagen und ihrer Übersetzungen in computerwissenschaftliche Methoden. In der wissenschaftlichen und philosophischen Literatur existieren zahlreiche Definitionen von Fairness. Diese reichen von „jeder bekommt das, was er verdient" (leistungsbezogen) über „alle bekommen das Gleiche" (egalitär) bis zu „jeder bekommt, was er braucht" (bedarfsorientiert) (vgl. Lamont/Favor 2008).

Diese Konzepte finden sich auch in der KI-Forschung wieder – jedoch in mathematisierter und operationalisierter Form: Dort stehen unterschiedliche statistische, kausale und individuelle Fairness-Definitionen nebeneinander, die auf jeweils verschiedene Dimensionen von Gleichbehandlung zielen. Beispielsweise operationalisiert „Demografische Parität" das Prinzip von egalitärer Fairness und „Chancengleichheit" das Prinzip der leistungsgezogenen Fairness. Verma und Rubin (2018) zeigen in ihrer vergleichenden Analyse, dass es keine universelle Fairness-Metrik gibt. Vielmehr stehen sich konkurrierende Ansätze gegenüber, die – je nach Anwendungsfall – unterschiedliche normative Prioritäten setzen und sich teilweise sogar gegenseitig ausschließen. Diese Vielfalt der Fairness-Konzepte stellt Organisationen vor die strategische Aufgabe, kontextabhängig zu entscheiden, welches Fairness-Ziel für ihre Anwendungen maßgeblich ist. Im Folgenden werden exemplarische Fairness-Konzepte und ihre mathematische Umsetzung vorgestellt, die für die Bewertung von KI-gestützten Entscheidungssystemen in Praxisfeldern wie Kreditvergabe, Personalgewinnung oder Leistungsbewertung besonders relevant sind.

3. Ausgewählte Fairness-Konzepte und ihre Umsetzung in Algorithmen

3.1 Demografische Parität (Demographic Parity)

Demografische Parität bedeutet, dass alle Gruppen – unabhängig von sensiblen Merkmalen wie Geschlecht, Alter oder Herkunft – mit der gleichen Wahrscheinlichkeit ein positives Ergebnis durch das KI-System erhalten. In einem Kreditbewertungssystem etwa hätten Männer und Frauen dieselbe Chance auf eine Kreditzusage, selbst wenn historische Daten ungleiche Ausgangsbedingungen widerspiegeln. Im HR-Kontext zeigt sich demografische Parität darin, dass algorithmisch ausgewählte Bewerber*innen zu gleichen Anteilen aus verschiedenen Alters- oder Geschlechtsgruppen stammen. Feldman et al. (2015) haben ein Verfahren entwickelt, um demografische Parität durch Pre-Processing herzustellen: Dabei werden Trainingsdaten so angepasst, dass indirekte Diskriminierungen über korrelierte Variablen reduziert werden. Als pragmatische Schwelle dient im US-amerikanischen Recht die sogenannte 80 %-Regel: Sie gilt als Indikator für potenziellen „Disparate Impact", wenn die Erfolgswahrscheinlichkeit einer Gruppe weniger als 80 % der Erfolgsrate der Vergleichsgruppe beträgt. Demografische Parität ist damit nicht nur ein technisches Kriterium, sondern hat auch rechtliche und gesellschaftspolitische Relevanz.

Relevanz für Manager*innen: Diese Definition ignoriert Unterschiede, wie z. B. in der Qualifikation von Bewerber*innen und kann damit zu Zielkonflikten mit leistungsbasierten Gerechtigkeitsprinzipien führen. Ihre Anwendung sollte deshalb bewusst abgewogen werden.

3.2 Chancengleichheit (Equal Opportunity)

Chancengleichheit bedeutet, dass Personen, für die ein positives Ergebnis sachlich gerechtfertigt ist – etwa weil ihr Kreditrisiko als tragbar eingeschätzt wird – unabhängig von ihrer Gruppenzugehörigkeit mit gleicher Wahrscheinlichkeit ein positives Outcome erhalten. In einem Kreditbewertungssystem heißt das: Antragsteller*innen mit vergleichbarem Rückzahlungsrisiko sollen – unabhängig von Geschlecht, Herkunft oder Alter – gleich häufig eine Kreditzusage erhalten. Hardt et al. (2016) haben dieses Fairnesskonzept für binäre Klassifikationssysteme eingeführt. Sie zeigen, dass sich Chancengleichheit durch ein nachgelagertes Post-Processing erreichen lässt: Dabei werden die Entscheidungsschwellen gruppenspezifisch angepasst, um vergleichbare Trefferquoten (True Positive Rates) für alle Gruppen sicherzustellen.

Relevanz für Manager*innen: Im Unterschied zu „Equalized Odds" („ausgeglichene Quoten") berücksichtigt Equal Opportunity nicht die Fehlalarme („False Positive Rates") – es ist daher ein weniger strenges, aber oft praktikableres Fairness-Kriterium.

3.3 Individuelle Fairness (Individual Fairness)

Individuelle Fairness folgt dem Prinzip: „Ähnliche Fälle sollen ähnlich behandelt werden". Zwei Personen mit vergleichbarer Ausbildung, Berufserfahrung und Leistung sollten – unabhängig von Geschlecht, Herkunft oder Alter – die gleiche Chance auf eine Beförderung erhalten. Technisch basiert dieser Ansatz auf der Definition geeigneter Ähnlichkeitsmetriken, anhand derer das System entscheiden kann, ob zwei Fälle vergleichbar sind (vgl. Dwork et al. 2012). Im Unterschied zu gruppenbezogenen Fairnessansätzen betrachtet individuelle Fairness nicht die statistische Gleichbehandlung ganzer Gruppen, sondern die gerechte Behandlung einzelner Personen im Kontext ihrer Merkmale.

Relevanz für Manager*innen: Individuelle Fairness ist besonders wichtig in Bereichen, in denen

personenbezogene Entscheidungen getroffen werden – etwa im HR-Management, bei Beförderungen, Leistungsbewertungen oder Schulungsangeboten. Unternehmen, die hier fair agieren wollen, müssen sicherstellen, dass vergleichbare Mitarbeitende auch vergleichbare Entwicklungschancen erhalten.

3.4 Prozedurale Fairness

Prozedurale Fairness bezieht sich nicht auf das Ergebnis einer Entscheidung, sondern auf den Weg dorthin – also darauf, wie nachvollziehbar, transparent und erklärbar ein algorithmischer Entscheidungsprozess ist. Wenn ein KI-System zum Beispiel entscheidet, welcher Kunde ein bestimmtes Angebot erhält oder welche Person in einem Auswahlverfahren weiterkommt, dann muss verständlich und begründbar sein, warum diese Entscheidung so getroffen wurde. Gerade bei komplexen KI-Modellen, etwa auf Basis von Deep Learning, besteht die sog. „Black-Box-Problematik": Die Entscheidung ist für Menschen – einschließlich Fachabteilungen, Kunden oder Aufsichtsbehörden – nicht mehr nachvollziehbar. Das erschwert die Überprüfung, verhindert Einspruchsmöglichkeiten und kann das Vertrauen in automatisierte Entscheidungen untergraben. Ansätze der „Explainable Artificial Intelligence" (XAI) setzen genau hier an. Sie verfolgen das Ziel, Entscheidungsprozesse in KI-Systemen sichtbar und interpretierbar zu machen. Verfahren wie LIME oder SHAP gehören zu den etablierten XAI-Methoden: Sie identifizieren und gewichten jene Merkmale, die in einem konkreten Fall ausschlaggebend für die Entscheidung waren (vgl. Arrieta et al. 2019). Allerdings hat XAI auch klare Grenzen. Die Erklärungen basieren meist auf vereinfachten, lokalen Annäherungen an das tatsächliche Modellverhalten und können komplexe Interaktionen nur eingeschränkt abbilden. Zudem besteht die Gefahr, dass vermeintlich plausible Erklärungen eine falsche Sicherheit vermitteln, ohne tatsächlich die Systemlogik offenzulegen. XAI ist damit nur ein Baustein für eine Fairness-Governance im Unternehmen.

Relevanz für Manager*innen: XAI gilt als vielversprechender Ansatz, um die Transparenzanforderungen des EU AI Act zu erfüllen. Gerade bei Hochrisiko-Systemen ist die Nachvollziehbarkeit der Entscheidungslogik keine Option, sondern eine regulatorische Pflicht – XAI kann hier helfen, Compliance sicherzustellen und Vertrauen bei Stakeholdern zu stärken.

Die Umsetzung von Fairness in algorithmischen Entscheidungssystemen ist komplex, weil die Hürden auf mehreren Ebenen liegen – von der theoretischen Definition bis zur praktischen

Anwendung. Schon auf konzeptioneller Ebene zeigt sich, dass Fairness nicht eindeutig bestimmbar ist: Unterschiedliche Fairness-Definitionen beruhen auf teils widersprüchlichen Annahmen und lassen sich mathematisch nicht gleichzeitig erfüllen. Hinzu kommen normative Grundentscheidungen, kulturelle Unterschiede und kontextspezifische Anforderungen, die bestimmen, was als „fair" gilt. Verschärft wird die Situation durch fehlende Standards und Benchmarks, die Unternehmen Orientierung geben könnten. Das folgende Kapitel beleuchtet diese Herausforderungen im Detail.

4. Herausforderungen bei der Umsetzung von Fairness in Algorithmen

4.1 Widersprüchliche Fairnesskonzepte

Es ist mathematisch nachgewiesen, dass nicht alle Fairness-Definitionen gleichzeitig erfüllbar sind – ein Sachverhalt, der als Fairness Impossibility Theorem bekannt ist (vgl. Verma/Rubin 2018). Entscheidet man sich beispielsweise für Chancengleichheit (Equal Opportunity), kann dies zu ungleichen Outcomes führen – etwa bei Gruppen mit unterschiedlichen Basisraten. Umgekehrt führt die Erfüllung von Outcome-Gleichheit (z. B. Demographic Parity) oft zu einer Ungleichbehandlung qualifizierter Personen. Unternehmen müssen deshalb Prioritäten setzen und begründen, welches Fairnesskriterium sie verfolgen.

4.2 Normative Grundentscheidungen

Jede Fairnessdefinition beruht auf einem bestimmten gesellschaftlichen Werteverständnis: Was als „fair" gilt, ist keine neutrale technische Entscheidung, sondern Ausdruck politischer, kultureller oder unternehmerischer Leitlinien. Das zeigt eindrücklich das Moral-Machine-Experiment des MIT (vgl. Awad et al. 2018), bei dem weltweit unterschiedliche moralische Entscheidungen in KI-gesteuerten Dilemmasituationen untersucht wurden. Die Ergebnisse zeigten: Fairnesspräferenzen variieren je nach Kulturraum, z. B. hinsichtlich Alter, Geschlecht oder gesellschaftlichem Status. Unternehmen müssen deshalb bewusst definieren, welches normative Fairnessverständnis sie vertreten – etwa im Rahmen ihrer CDR-Strategie oder gegenüber globalen Nutzergruppen.

4.3 Kontextabhängigkeit

Fairness ist nicht universell definierbar – selbst bei gleichen ethischen Grundannahmen können sich je nach Anwendung unterschiedliche Fairnessanforderungen ergeben. Was im HR-Bereich als gerecht gilt – etwa gleiche Chancen auf Bewerbung oder Beförderung – kann in anderen Bereichen wie Marketing, Gesundheitswesen oder Justiz anders bewertet werden. So kann etwa das gezielte Ansprechen bestimmter Nutzergruppen (z. B. Präventionskampagnen für Risikogruppen) als gerechtfertigte Differenzierung gelten, während dies im Recruiting diskriminierend wäre (vgl. Nepomuceno/Petrillo 2025). Fairnessmetriken wie Demographic Parity, Equal Opportunity oder Equalized Odds lassen sich daher nicht pauschal anwenden, sondern müssen jeweils auf Ziel, Branche und Wirkung des Systems abgestimmt werden. Dies erfordert von Unternehmen die Fähigkeit, Fairness als situativ auslegbares Prinzip zu denken – nicht als statische Compliance-Kennzahl.

4.4 Fehlende Standards und Benchmarks

Bislang existieren keine verbindlichen regulatorischen Standards für Fairness in KI-Systemen. Unternehmen stehen oft allein vor Fragen wie: Welche Fairness-Metrik ist angemessen? Welche Abweichungen gelten als tolerierbar? Wie kann Fairness überprüft, dokumentiert und auditiert werden? Frameworks wie AI Fairness 360 (vgl. Bellamy et al. 2019; Linux Foundation 2024) oder ISO/IEC 42001 „Artificial intelligence management system“ (vgl. ISO/IEC 2023) bieten erste Orientierung, ersetzen aber keine rechtlich verbindlichen Anforderungen. Hier besteht ein hoher Bedarf an Klarheit – sowohl für die strategische Steuerung als auch für Rechenschaftspflichten gegenüber Kund*innen, Öffentlichkeit und Aufsicht.

Diese Übersicht macht deutlich: Für Fairness in KI gibt es aktuell kein „Rezeptbuch“, keine Anwendungsroutine für Compliance. Es ist ein strategischer Prozess: Entscheider*innen müssen sich mit konkurrierenden Fairnessansätzen auseinandersetzen und den passenden für ihren Kontext wählen. Sie müssen die damit verbundenen Zielkonflikte aktiv managen – zwischen Effizienz und Ethik, zwischen Datenschutz und Differenzierung, zwischen unternehmerischem Erfolg und gesellschaftlicher Verantwortung. Und sie müssen bereit sein, Verantwortung zu übernehmen – auch für Unsicherheiten, Grauzonen und Dilemmata, die sich nicht auflösen lassen. Fairness ist kein Zustand, sondern eine Zumutung – aber eine notwendige. Wer sie aktiv gestaltet, handelt nicht nur verantwortungsvoll, sondern zukunftsorientiert.

5. Unsichtbare Ungleichheit – Wie Daten Fairness untergraben können

5.1 Wie Bias in die Systeme gelangt

„Die Daten lügen nicht", heißt es oft. Doch sie erzählen auch nicht die ganze Wahrheit. In ihnen steckt nicht nur Information, sondern Geschichte – eine Geschichte von Sichtbarkeit, Macht und Marginalisierung. Algorithmen lernen aus der Vergangenheit. Und die Vergangenheit war nicht neutral. Ein Bias – also eine Verzerrung – kann an vielen Stellen in die KI-Wertschöpfungskette gelangen (vgl. Barocas et al. 2023):

- Datenbasis: Trainingsdaten bilden vergangene Realität ab – inklusive gesellschaftlicher Ungleichheiten.
- Labeling: Menschliche Urteile beim Annotieren von Daten tragen unbewusste Vorurteile in Modelle.
- Feature-Auswahl: Merkmale wie Wohnort oder Sprachstil können unbeabsichtigt als Stellvertreter („Proxy-Variablen") für sensible Eigenschaften wirken.
- Systemarchitektur: Modelle, die rein auf maximale Vorhersagekraft optimiert sind, übernehmen bestehende Muster ungeprüft weiter.

5.2 Unsichtbare Reproduktion von Ungleichheit

Der gefährlichste Bias ist oft der, den niemand sieht. Wenn ein Recruiting-Algorithmus Lebensläufe bevorzugt, die bestimmte Muster enthalten (z. B. männlich konnotierte Formulierungen oder spezifische Studiengänge), dann entsteht Diskriminierung, ohne dass das System explizit nach Geschlecht fragt. Diese Effekte sind schwer zu erkennen – und noch schwerer zu beheben. Ein prominentes Beispiel: Amazon entwickelte ein KI-gestütztes Bewerbungssystem, das systematisch männliche Bewerber bevorzugte. Der Grund: Die Trainingsdaten basierten auf früheren Bewerbungen, bei denen männliche Profile überrepräsentiert waren. Das System lernte diese „Erfolgsmuster" – und verstärkte damit eine bereits bestehende Schieflage (vgl. Dastin 2018). Weitere Beispiele zeigen ähnliche Muster:

- Spracherkennungssoftware versteht weibliche Stimmen schlechter (vgl. Bajorek 2019).
- Werbealgorithmen zeigen hochbezahlte Jobs häufiger Männern (vgl. Lambrecht/Tucker 2019).
- Kredit-Scoring-Systeme benachteiligen Menschen aus einkommensschwachen Stadtteilen (vgl. Barocas/Selbst 2016).

5.3 Bias ist nicht böse – sondern strukturell

Wichtig ist: Bias ist selten das Ergebnis böswilliger Absicht. Vielmehr ist er Ausdruck struktureller Ungleichheiten, die sich in digitalen Systemen spiegeln – und dabei oft noch verstärkt werden. Viele KI-Systeme basieren auf historischen Trainingsdaten, die strukturelle Ungleichheiten widerspiegeln – etwa durch diskriminierende Praktiken, unausgewogene Erfassungsraten oder bestehende gesellschaftliche Vorurteile („Social Bias"). Diese Form des sozialen oder historischen Bias ist besonders kritisch, da sie oft unsichtbar, aber tief in der Datenbasis verankert ist.

Barocas und Selbst (2016) zeigen, dass Diskriminierung nicht erst im Modell entsteht, sondern häufig bereits in der Art, wie Daten erhoben, ausgewählt und genutzt werden. Verzerrungen können sich in der Datenauswahl, Feature-Selektion oder durch einseitige Zieldefinitionen manifestieren – und werden durch Modellierungsschritte oftmals noch verstärkt. Technische Korrekturen entlang des Machine-Learning-Zyklus (z. B. Pre-, In- oder Post-Processing) sind möglich, aber nie neutral: Auch sie beruhen auf normativen Entscheidungen darüber, welche Form von Ungleichheit als problematisch und welche als tolerierbar gilt (vgl. Suresh/Guttag 2021). Jede technische Intervention ist daher eingebettet in kontextuelle Wertentscheidungen, die sowohl die Art des erkannten Schadens als auch die gewählte Abhilfestrategie prägen.

Dieser Abschnitt verdeutlicht: Fairness in KI ist nicht nur eine Frage mathematischer Definitionen, sondern auch eine Frage der Datenethik, der sozialen Verantwortung und des kritischen Umgangs mit historischen Mustern.

5.4 Blinder Fleck: Intersektionalität

Doch selbst wenn diese Herausforderungen adressiert werden – durch technische Korrekturen, normativ reflektierte Modellierung oder sorgfältige Datenstrategien – bleibt ein weiterer blinder Fleck bestehen: Die meisten KI-Systeme denken entlang einzelner, isolierter Kategorien wie „Geschlecht", „Herkunft" oder „Alter". Doch Menschen sind komplex – und ihre Benachteiligungen oft das Ergebnis überlappender Diskriminierungen. Genau an diesen Schnittstellen versagt herkömmliche algorithmische Fairness.

„Intersektionalität" beschreibt die wechselseitige Verstärkung mehrerer Diskriminierungsformen – etwa wenn sich Geschlecht, Herkunft und Alter überschneiden. Der Begriff geht auf

die Juristin Kimberlé Crenshaw (1989) zurück und hat längst Eingang in die KI-Debatte gefunden. So zeigen Studien wie Buolamwini und Gebru (2018), dass Gesichtserkennungssysteme besonders häufig bei dunkelhäutigen Frauen versagen – nicht, weil das System absichtlich diskriminiert, sondern weil diese spezifische Gruppenkonstellation in den Trainingsdaten kaum vertreten war.

Die Forschung macht deutlich: Intersektionale Fairness ist dringend notwendig – aber schwer umzusetzen. Sie erfordert granulare Daten, komplexe Metriken und vor allem ein neues Bewusstsein für Mehrfachdiskriminierung. Für die Praxis bedeutet das: Auch wenn praktikable Lösungen (noch) fehlen, darf das Thema nicht ignoriert werden. Wer Verantwortung für faire KI übernehmen will, muss intersektionale Perspektiven mitdenken – und technologische wie regulatorische Entwicklungen zur Operationalisierung aufmerksam verfolgen.

Fairness in der algorithmischen Entscheidungsfindung ist kein rein technisches oder rechtliches Thema. Sie ist ein normativ aufgeladenes Feld, in dem unterschiedliche Interessen und Werte aufeinandertreffen. In der unternehmerischen Praxis ergeben sich daraus sowohl Konflikte zwischen legitimen Zielen als auch ethische Dilemmata, bei denen keine Lösung aus ethischer Sicht vollständig zufriedenstellen ist. Das folgende Kapitel beleuchtet diese Spannungsfelder, mit denen Entscheider*innen in der Gestaltung fairer KI-Systeme konfrontiert sind – und liefert Reflexionsfragen für eine verantwortungsvolle Navigation.

6. Ethische Dilemmata und Zielkonflikte – Warum Fairness Mut braucht

6.1 Neutralitätsillusion vs. verzerrte Datenrealität

Viele KI-Systeme beruhen auf historischen Daten, die bestehende Ungleichheiten widerspiegeln. Diese Verzerrungen entstehen nicht durch böse Absicht, sondern durch die Struktur der Realität: In Bewerbungssystemen zeigen sich Vorurteile gegenüber bestimmten Hochschulen oder Sprachstilen; in Kreditsystemen historische Benachteiligung bestimmter Wohngegenden. Werden diese Muster unreflektiert fortgeschrieben, entsteht algorithmische Diskriminierung.

6.2 Fairness vs. Genauigkeit und Effizienz

Fairness hat ihren Preis – zumindest kurzfristig. Denn die Umsetzung von Fairness-Kriterien in KI-Systemen kann die statistische Genauigkeit leicht verringern. In der Praxis bedeutet das: Ein Modell, das darauf optimiert ist, alle Gruppen gerecht zu behandeln, kann insgesamt etwas weniger präzise vorhersagen. Ein typisches Beispiel ist der Zielkonflikt zwischen Chancengleichheit (Equal Opportunity) und der Gesamtgüte eines Modells, etwa gemessen an der „Receiver Operating Characteristic – Area Under the Curve" ROC-AUC. In der Unternehmenspraxis führt das zu einer strategischen Abwägung: Wollen wir ein System, das gerechter gegenüber verschiedenen Gruppen ist – auch wenn es etwas weniger treffsicher oder effizient ist? Oder priorisieren wir maximale Genauigkeit und Geschwindigkeit, um operative Anforderungen oder KPIs zu erfüllen?

6.3 Datenschutz vs. Personalisierung

Damit KI-Systeme faire Entscheidungen im Einzelfall treffen können, brauchen sie oft detaillierte Informationen über einzelne Personen – etwa zu Bildung, Herkunft, Sprache oder Verhalten. Je mehr ein System über eine Person weiß, desto gezielter kann es auf deren Situation reagieren. Doch genau darin liegt das Dilemma: Mehr Personalisierung erhöht das Risiko, dass sensible Daten missbraucht, falsch interpretiert oder zu anderen Zwecken verwendet werden. Unternehmen laufen Gefahr, Grenzen des Datenschutzes zu überschreiten oder unbeabsichtigt diskriminierende Profile zu erzeugen („Overprofiling"). Die zentrale Frage lautet also: „Wie viel Wissen über Menschen ist notwendig – und wie viel ist vertretbar?"

6.4 Ergebnisgerechtigkeit vs. Verfahrensgerechtigkeit

Fairness kann sich auf das Was (das Ergebnis) oder das Wie (den Prozess) einer Entscheidung beziehen – idealerweise auf beides. Doch in der Praxis stehen diese Ansprüche oft in Spannung zueinander. Ein KI-System kann formal korrekt und gut dokumentiert sein, aber systematisch benachteiligen – etwa durch die Nutzung von Variablen, die für bestimmte Gruppen nachteilig wirken. Ebenso kann ein System ausgleichende Ergebnisse liefern – z. B. gleiche Chancen für verschiedene Gruppen – ohne dass der Entscheidungsweg transparent oder erklärbar ist. Diese Reflexionsfragen

müssen sich Entscheider*innen stellen: Reicht es uns, dass der Entscheidungsprozess dokumentiert und nachvollziehbar ist – selbst wenn das Ergebnis bestimmte Gruppen benachteiligt? Oder priorisieren wir gerechte Resultate – auch wenn wir die Entscheidungen nicht vollständig erklären können? Wie kommunizieren wir diese Entscheidungen gegenüber Betroffenen, Aufsichtsbehörden oder Öffentlichkeit? Welche Art von Vertrauen ist uns wichtiger – Vertrauen in den Prozess oder in das Ergebnis? In der Praxis hilft eine Balance: Transparente Verfahren schaffen Vertrauen. Gerechte Ergebnisse sichern Akzeptanz. Wer Fairness gestalten will, muss beide Seiten bewusst austarieren.

6.5 Individuelle Fairness vs. gruppenbezogene Fairness

Fairness kann sich entweder an Gruppen oder am Einzelfall orientieren – und genau darin liegt ein häufig übersehener Zielkonflikt. Gruppenbezogene Ansätze – etwa Demographic Parity – zielen darauf, dass z. B. Frauen und Männer vergleichbare Chancen auf ein positives Ergebnis haben. Das hilft, strukturelle Benachteiligung sichtbar zu machen und aktiv auszugleichen. Individuelle Fairness dagegen fragt: Werden vergleichbare Personen unabhängig von ihrer Gruppenzugehörigkeit gleichbehandelt? In der Praxis entsteht hier ein Spannungsfeld: Eine Quote kann im Einzelfall als ungerecht erscheinen, wenn leistungsstärkere Kandidat*innen übergangen werden. Umgekehrt kann eine rein individuelle Auswahl bestehende Ungleichheiten zementieren – etwa, weil Menschen mit weniger Ressourcen seltener als „Top-Performer" gelten. Für Entscheider*innen stellt sich deshalb die Frage: Wie balancieren wir Chancengleichheit und Leistungsgerechtigkeit – und welche Fairness wollen wir für einen konkreten Anwendungsfall konkret gestalten?

6.6 Fairness vs. bestehende Anreizsysteme

Fairnessziele stehen oft im Spannungsfeld zu etablierten Leistungskennzahlen. Wenn Abteilungen zum Beispiel auf schnelle Einstellungen („Time to Hire") oder rein auf Spitzenleistung („Best Qualified") optimieren, kann eine Maßnahme zur Förderung von Diversität als hinderlich empfunden werden. Auch in Vertrieb oder Marketing folgen Boni-Modelle häufig rein ökonomischen Logiken – nicht der fairen Repräsentanz. Entscheider*innen müssen daher abwägen: Wie lassen sich Fairnessziele in bestehende Anreizsysteme integrieren – und wo braucht es vielleicht ein Umdenken bei KPIs und Erfolgsdefinitionen?

6.7 Kurzfristige Logiken vs. intergenerationelle Fairness

Fairness ist nicht nur eine Frage des Jetzt. Entscheidungen in KI-Systemen, etwa in Bildungswegen, Kreditbewilligungen oder Matching-Prozessen, haben oft langfristige Effekte. Eine auf kurzfristige KPI-Optimierung ausgerichtete Logik vernachlässigt diese Langzeitwirkungen. Intergenerationelle Fairness fragt: Wie verhindern wir, dass sich heute marginalisierte Gruppen auch morgen im Nachteil befinden? Fairness bedeutet nicht, Zielkonflikte zu vermeiden – sondern sie anzuerkennen, transparent zu machen und im Sinne von Unternehmenswerten auszubalancieren. Das braucht Mut zur Ambiguität, Lust auf Aushandlung und eine klare ethische Positionierung. Nur dann wird Fairness zur gestaltbaren Ressource in einer algorithmisch erweiterten Unternehmenswelt.

Fairness kann durch bewusste Entscheidungen im Management entlang der gesamten Entwicklung und Nutzung von KI-Systeme entstehen. Die dargestellten Zielkonflikte – zwischen Fairness und Effizienz, Datenschutz, Transparenz oder Gleichbehandlung – zeigen: Es gibt keine einfache Lösung. Doch es gibt Gestaltungsräume. Das nachfolgende Kapitel bietet einen praxisorientierten Leitfaden, wie Organisationen Fairness im KI-Einsatz systematisch verankern können – von der Zieldefinition bis zur Wirkungskontrolle.

7. Fairness gestalten – Entscheidungen entlang des KI-Lebenszyklus

7.1 Ausgangslage und zentrale Problemfelder

Die Umsetzung von Fairness in algorithmischen Entscheidungssystemen ist komplex, weil die Hürden auf mehreren Ebenen liegen – von der theoretischen Definition bis zur praktischen Anwendung. Schon auf konzeptioneller Ebene zeigt sich, dass Fairness nicht eindeutig bestimmbar ist: Unterschiedliche Fairness-Definitionen beruhen auf teils widersprüchlichen Annahmen und lassen sich mathematisch nicht gleichzeitig erfüllen. Hinzu kommen normative Grundentscheidungen, kulturelle Unterschiede und kontextspezifische Anforderungen, die bestimmen, was als „fair" gilt. Verschärft wird die Situation durch fehlende Standards und Benchmarks, die Unternehmen Orientierung geben könnten. Die folgenden Abschnitte beleuchten diese Herausforderungen im Detail.

7.2 Anforderungen: Fairness früh mitdenken

Ein zentrales Dilemma bei der Entwicklung fairer KI-Systeme besteht im Spannungsfeld zwischen reaktiver Reparatur („Fairness-by-Debugging") und proaktiver Gestaltung („Fairness-by-Design"). Viele Probleme in der Fairness entstehen, weil sie zu spät bedacht wird – erst wenn das System bereits läuft und Verzerrungen sichtbar werden. Dieses nachträgliche Eingreifen nennt man Fairness-by-Debugging. Dabei werden bestehende Ungleichheiten im Nachhinein korrigiert – oft mit hohem Aufwand und begrenzter Wirkung. Deutlich wirkungsvoller ist Fairness-by-Design: Hier wird Fairness von Anfang in die Zieldefinition, die Auswahl der Daten und die Modellgestaltung integriert. Wer schon in der Konzeptionsphase fragt, welche Wirkung ein KI-System auf welche Gruppen hat, kann viele Zielkonflikte vermeiden. Praxisansätze sind:

- Früher Stakeholder-Dialog zur Zielklärung
- Bewusste Wahl eines Fairnesskonzepts (z. B. Chancengleichheit oder demografische Parität)
- Ethik-Folgenabschätzung vor Projektstart, um Risiken und Nebenwirkungen frühzeitig zu erkennen

7.3 Datenmanagement: Verzerrungen erkennen und vermeiden

Ein häufiges Dilemma im Datenmanagement ergibt sich aus dem Spannungsfeld zwischen historischer Verzerrung und Repräsentativität. Daten sind nie neutral – sie spiegeln vergangene Machtverhältnisse und machen diese in technischen Systemen nutzbar. Wer Fairness ernst nimmt, muss seine Trainingsdaten nicht nur technisch prüfen, sondern auch sozial und normativ hinterfragen. Gleichzeitig kann der Versuch, alle Verzerrungen zu eliminieren, zu Datenlücken oder Datenschutzproblemen führen. Praxisansätze sind:

- Audits der algorithmischen Systeme, z. B. in den Phasen Rahmenbestimmung, Strukturerfassung, Artefaktsammlung, Testen und Reflexion (vgl. Raji et al. 2020).
- Rebalancing-Verfahren (z. B. Synthetic Minority Over-sampling Technique SMOTE), welche synthetische Datenpunkte für unterrepräsentierte Gruppen erzeugen.
- Datasheets for Datasets fördern Transparenz und helfen, Verzerrungen bereits bei der Auswahl und Dokumentation von Trainingsdaten sichtbar zu machen (vgl. Gebru et al. 2021).

7.4 Modellentwicklung: Fairness und Performance austarieren

In der Modellentwicklung zeigt sich ein zentrales Dilemma: der Zielkonflikt zwischen Gerechtigkeit und Genauigkeit. Ein verbreitetes Missverständnis ist, Fairness ließe sich einfach als zusätzliche Variable hinzufügen. In Wahrheit verändert jede Fairnessvorgabe – etwa gleiche Fehlerraten – die Modelllogik und kann zu Zielkonflikten führen. Etwa, wenn die Modellgüte sinkt oder mehr False Positives entstehen. Manager*innen müssen hier bewusst abwägen: Ist höchste Präzision das oberste Ziel – oder gerechtere Behandlung bei leichtem Effizienzverlust? Praxisansätze sind:

- Auswahl passender Fairness-Kennzahlen (z. B. gleiche Fehlerquoten für verschiedene Gruppen)
- Einsatz von Software-Tools wie Fairlearn oder AI Fairness 360
- Testszenarien, um Fairness schon vor dem Einsatz eines KI-Systems zu prüfen

7.5 Betrieb und Monitoring: Transparenz und Kontrolle sicherstellen

Im laufenden Betrieb zeigt sich ein weiteres Dilemma: der Zielkonflikt zwischen prozeduraler Fairness und Ergebnisfairness. Selbst ein erklärbares Modell kann systematisch ungerechte Entscheidungen treffen – und ein intransparentes Modell kann zufällig faire Resultate liefern. Deshalb braucht es beides: erklärbare Entscheidungen und laufende Kontrolle der Auswirkungen. Wichtig ist zudem die Kommunikation nach außen: Fairness entsteht auch durch Vertrauen. Praxisansätze sind:

- Einsatz von XAI (z. B. SHAP, LIME)
- Einrichtung von Fairness-Dashboards
- Nutzung von Model Cards mit übersichtlichen Informationen zu Trainingsdaten, Einsatzgrenzen und Risiken eines KI-Modells (vgl. Mitchell et al. 2019)
- Regelmäßige Audits (intern/extern)

7.6 Wirkungskontrolle: Fairness im Zeitverlauf denken

In der Wirkungskontrolle prallen oft zwei Perspektiven aufeinander: die kurzfristige Optimierung aktueller Kennzahlen und die Sicherung langfristiger Fairness. Viele Effekte algorithmischer Entscheidungen werden erst mit zeitlichem Abstand sichtbar. Diskriminierende Auswahlprozesse im Recruiting formen den Führungskräftepool von morgen, und ein rein gegenwartsorientiertes Kredit-Scoring kann bestehende ökonomische Ausgrenzung verfestigen. Intergenerationelle Fairness

rückt daher die Frage in den Mittelpunkt, welche Folgen heutige Entscheidungslogiken für kommende Generationen haben. Praxisansätze sind:

- Impact-Assessments mit Langfristperspektive (z. B. SustAIn, vgl. Rohde et al. 2021)
- Einbindung von Perspektiven marginalisierter Gruppen

Fairness in KI ist keine rein technische Herausforderung. Sie ist eine strategische Gestaltungsaufgabe – und damit untrennbar mit Führung verbunden. Denn wer algorithmische Systeme einführt, entscheidet nicht nur über Funktionalität, sondern über Teilhabe, Gerechtigkeit und Zukunftsfähigkeit. CDR unterstützt Führungskräfte dabei, diesen Verantwortungsspielraum reflektiert und richtungsweisend auszufüllen.

8. Fairness führen – CDR als strategischer Kompass

8.1 Fairness als Führungsaufgabe

Künstliche Intelligenz verändert nicht nur Prozesse – sie verändert Machtverhältnisse. Wer heute algorithmische Systeme einführt, gestaltet damit neue Entscheidungspfade, beeinflusst Sichtbarkeit, Teilhabe und Gerechtigkeit. Und mehr noch: Er oder sie prägt damit das Verhältnis zwischen Organisation, Technologie und Gesellschaft. In dieser neuen Verantwortungskonstellation reicht technologische Exzellenz allein nicht aus. Es braucht Haltung, ethische Orientierung und strategische Steuerung. Genau hier setzt Corporate Digital Responsibility (CDR) an: als ethisch-normativer Ordnungsrahmen für digitale Transformationsprozesse – und als praktischer Kompass für faire KI-Systeme (vgl. Dörr 2021; Elliot et al. 2021).

8.2 Fairness beginnt nicht im Code, sondern im Kopf der Führung

CDR versteht sich nicht als technisches Regelwerk, sondern als Organisations- und Führungsprinzip. Sie verbindet Werte wie Transparenz, Teilhabe und Rechenschaft mit organisationaler Verantwortung. Fairness ist dabei keine Zusatzaufgabe für Tech-Abteilungen, sondern ein zentraler Bestandteil unternehmerischer Integrität – verankert in Governance-Strukturen, Zielsystemen, Leadership-Trainings und Wertedialogen (vgl. Dörr 2025; Herden et al. 2021; Lobschat et al. 2021). Nur wenn Fairness Teil der Führungskultur ist, kann sie entlang des gesamten digitalen

Ökosystems wirksam werden – von der Datenerhebung bis zur Entscheidungsfindung, von internen Prozessen bis zu gesellschaftlichen Auswirkungen (vgl. Kunz/Wirtz 2023). Fairness ist keine technische Stellschraube für KI, die sich nachjustieren lässt. Sie ist ein Ausdruck unternehmerischer Reife und Zukunftsfähigkeit. Unternehmen, die CDR ernst nehmen, betreiben keine „tick-box compliance" (vgl. Elliott et al. 2021), sondern entwickeln Führung neu – als reflexive, wertegeleitete Praxis, die mit Unsicherheit, Zielkonflikten und Ambiguitäten umgehen kann. Fairness in KI bedeutet, Verantwortung zu übernehmen – auch dort, wo sich normative Dilemmata nicht auflösen lassen. CDR macht diese Ambivalenzen sichtbar, ohne sie zu nivellieren. Sie fördert eine Haltung, die Position bezieht, zuhört und aushält. Und genau das ist der Kern zukunftsorientierter Führung.

8.3 Human-Centered Digital Leadership – Verantwortung für das Unsichtbare übernehmen

Ein zukunftsweisender Ansatz für digitale Führung lautet: Human-Centered Digital Leadership (vgl. Flink et al. 2024). Er erweitert die traditionelle Logik des „Digital Leadership" um eine menschenzentrierte Perspektive und verbindet wirtschaftliche Ziele mit sozialen Auswirkungen.

> Eine auf den Menschen ausgerichtete Führung schafft ein Gleichgewicht zwischen finanzieller Wertschöpfung und den Auswirkungen auf alle Stakeholder, die durch digitale Technologien betroffen sind – wirtschaftlich, psychologisch, ökologisch oder rechtlich (Flink et al. 2024: 2).

Dieser Ansatz sensibilisiert Führungskräfte für das, was oft übersehen wird: die unbeabsichtigten, aber realen Konsequenzen von nicht auf Fairness ausgerichteten KI-Systemen. Zu den unsichtbaren Folgen digitaler Technologien zählen beispielsweise:

- versteckte Diskriminierungen durch algorithmische Voreingenommenheit,
- digitale Exklusion durch Inkompatibilität mit Randgruppen,
- neue Machtasymmetrien durch Datenmonopole.
- Digital Sensing (vgl. Flink et al. 2024) bietet konkrete Reflexionsräume und Führungspraktiken, um diese Effekte proaktiv zu adressieren und KI- Strategien nicht nur an Marktchancen, sondern auch an gesellschaftlicher Wirkung auszurichten
- Bewusstsein schaffen: Was sehen wir (noch) nicht?
- Stakeholder integrieren: Wer ist betroffen, aber nicht beteiligt?

- ESG-Ziele mitdenken: Wie trägt Technologie zu Gerechtigkeit, Umwelt und Governance bei?
- Menschzentrierte Entscheidungen treffen: Was bedeutet unser Handeln für die Schwächsten im System?

Human-Centered Leadership fordert, diese „unsichtbaren Folgen" systematisch zu identifizieren, zu bewerten und zu managen. Damit ist sie direkt anschlussfähig an die CDR-Perspektive, die Fairness als kontinuierlichen Reflexionsprozess versteht.

8.4 Führungsverantwortung sichtbar machen: Fairness in der Entscheidungspraxis verankern

Fairness entsteht durch Entscheidungen – insbesondere dann, wenn digitale Systeme über Zugang, Chancen und Teilhabe mitentscheiden. Deshalb braucht es eine neue Führungsperspektive: Human Impact Decision Making, d.h. den menschlichen Wirkungsradius von Entscheidungen systematisch mitzudenken – bevor, nicht erst nachdem KI-Systeme wirken. Dieser Ansatz, entwickelt im Rahmen der Human-Centered Digital Leadership (Flink et al. 2024), fordert, dass unternehmerische Entscheidungen auch an ihren Auswirkungen auf Menschen und Gesellschaft gemessen werden – nicht nur an Effizienz und KPIs. Konkret bedeutet das für die Fairness im Umgang mit KI-Systemen (vgl. Abbildung 2):

- Kodifizierung bestehender Prinzipien: Gibt es bereits definierte Fairnessziele, etwa im Rahmen einer CDR-Strategie, in Ethik-Leitlinien für KI oder in projektbezogenen Selbstverpflichtungen? Wird dokumentiert, welche Fairnessdefinition (z. B. Chancengleichheit vs. demografische Parität) im jeweiligen Anwendungsfall verfolgt wird – etwa im Recruiting oder im Kundenservice?
- Schutzbedürftige priorisieren: Wird systematisch analysiert, welche sozialen Gruppen potenziell in KI-Anwendungen benachteiligt werden könnten – etwa aufgrund von Herkunft, Alter, Sprache oder sozioökonomischem Hintergrund? Werden intersektionale Diskriminierungsrisiken bedacht – also das Zusammenspiel mehrerer benachteiligender Faktoren. Und: Sind betroffene Gruppen in die Entwicklung und Bewertung von KI-Systemen eingebunden?
- Potenzielle Nebenwirkungen sichtbar machen: Welche impliziten Annahmen liegen Trainingsdaten oder Zielmetriken zugrunde – und welche sozialen Verzerrungen könnten dadurch fortgeschrieben werden? Werden mögliche „Nebenwirkungen" wie mangelnde

Transparenz, Ausschlusswirkungen oder Reputationsrisiken frühzeitig reflektiert – z. B. durch Impact-Assessments oder Bias-Audits?

- Vertrauen schaffen: Wie kommuniziert das Unternehmen seine Prinzipien, Prüfprozesse und Entscheidungen rund um algorithmische Fairness nach innen und außen? Gibt es erklärbare Modelle, zugängliche Dokumentation, Beschwerdemechanismen? Ist die Entscheidungslogik – etwa bei automatisierten Bewertungen oder Empfehlungen – für Stakeholder nachvollziehbar und revisionsfähig?

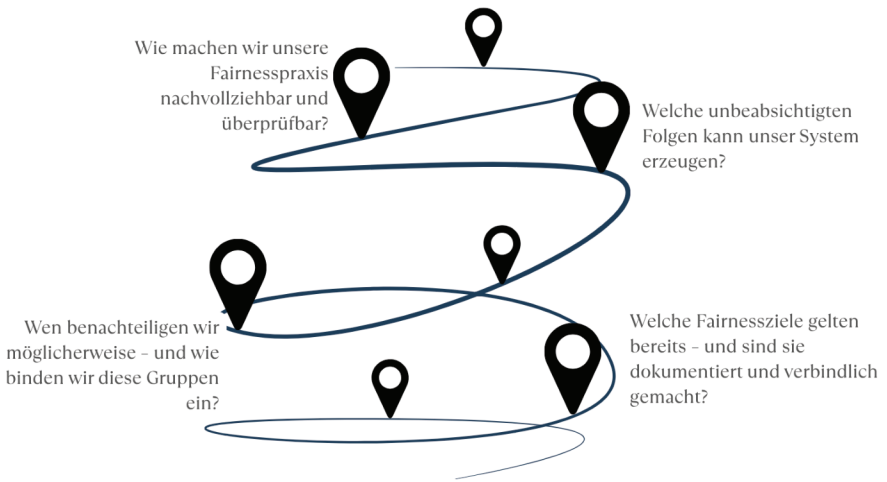

ABBILDUNG 2: FÜHRUNGSFRAGEN FÜR FAIRNESS IN KI-SYSTEMEN (OPERATIONALISIERUNG VON HUMAN IMPACT DECISION MAKING) (QUELLE: EIGENE ABBILDUNG)

Diese Dimensionen bilden einen Orientierungsrahmen für faire Führung im digitalen Kontext – und sind zugleich eine konkrete Umsetzung der Prinzipien der CDR (vgl. Dörr 2025: 208–222). Sie operationalisieren Fairness als unternehmerisches Führungsversprechen, das Wirkung entfaltet – nach innen wie außen, kurzfristig wie langfristig. Dabei geht es nicht um moralische Makellosigkeit, sondern um offene, begründete, inklusive Entscheidungen, die an ethischen Maßstäben und gesellschaftlicher Wirkung ausgerichtet sind (vgl. Elliott et al. 2021).

Die digitale Transformation verlangt mehr als Innovation und Effizienz. Sie verlangt ethische Navigation – durch komplexe Datenräume, durch normative Zielkonflikte, durch neue Verantwortungsbeziehungen zwischen Mensch, Maschine und Organisation. Corporate Digital Responsibility (CDR) bietet dafür den notwendigen Kompass. Sie verbindet technische Gestaltung mit ethischer Reflexion, Governance mit Haltung, Compliance mit Kultur. Sie macht sichtbar: Es genügt nicht, dass Algorithmen funktionieren – sie müssen auch gerecht wirken.

Fairness ist dabei kein messbarer Endzustand. Sie ist ein kontinuierlicher Führungsakt – ein Balanceakt zwischen Ambiguitäten, Zielkonflikten und Wertespannungen. Ein Akt, der beginnt, wenn Führungsteams bereit sind, Verantwortung zu übernehmen: Für das Sichtbare und das Unsichtbare, für die heutigen Entscheidungen über KI-Anwendungen und deren zukünftige Wirkungen, für Menschen mit und für die sie tätig sind. Entscheider*innen, die Fairness als Führungsaufgabe begreifen, handeln nicht nur regelkonform und ethisch. Sie gestalten die Voraussetzungen für eine gerechtere digitale Zukunft.

Literaturverzeichnis

Arrieta, A. B. / Díaz-Rodríguez, N. / Del Ser, J. / Bennetot, A. / Tabik, S. / Barbado, A. / García, S. / Gil-López, S. / Molina, D. / Benjamins, R. / Chatila, R. / Herrera, F. (2019): Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI, in: Information Fusion, Jg. 58 / Nr. C, 82–115, DOI: 10.1016/j.inffus.2019.12.012..

Awad, E. / Dsouza, S. / Kim, R. / Schulz, J. / Henrich, J. / Shariff, A. / Bonnefon, J.-F. / Rahwan, I. (2018): The Moral Machine Experiment, in: Nature, Jg. 563 / Nr. 7729, 59–64, DOI: 10.1038/s41586-018-0637-6 (aufgerufen am: 21/06/2025).

Bajorek, J. P. (2019): Voice Recognition Still Has Significant Race and Gender Biases, URL: https://hbr.org/2019/05/voice-recognition-still-has-significant-race-and-gender-biases (aufgerufen am: 14/06/2025).

Barocas, S. / Hardt, M., / Narayanan, A. (2023). Fairness and Machine Learning: Limitations and Opportunities, URL: https://fairmlbook.org (aufgerufen am: 04/08/2025).

Barocas, S. / Selbst, A. D. (2016): Big Data's Disparate Impact, in: California Law Review, Jg. 104 / Nr. 3, 671–732, DOI: 10.2139/ssrn.2477899 (aufgerufen am: 14/06/2025).

Bellamy, R. K. E. / Dey, K. / Hind, M. / Hoffman, S. C. / Houde, S. / Kannan, K. / Lohia, P. / Martino, J. / Mehta, S. / Mojsilovic, A. / Nagar, S. / Natesan Ramamurthy, K. / Richards, J. / Saha, D. / Sattigeri, P. / Singh, M. / Varshney, K. / Zhang, Y. (2019): AI Fairness 360: An Extensible Toolkit for Detecting and Mitigating Algorithmic Bias, in: IBM Journal of Research and Development, Jg. 63 / Nr. 4 / 5, 1–15, DOI: 10.1147/JRD.2019.2942287 (aufgerufen am: 21/06/2025).

Buolamwini, J. / Gebru, T. (2018): Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification, in: Friedler, S. A. / Wilson, C. (Hrsg.): Proceedings of the 1st Conference on Fairness, Accountability and Transparency (FAT*), 77–91, URL: https://proceedings.mlr.press/v81/buolamwini18a.html (aufgerufen am: 14/06/2025).

Crenshaw, K. (1989): Demarginalizing the Intersection of Race and Sex: A Black Feminist Critique of Antidiscrimination Doctrine, Feminist Theory and Antiracist Politics, in: University of Chicago Legal Forum, Jg. 1989 / Nr. 1, Artikel 8, URL: https://chicagounbound.uchicago.edu/uclf/vol1989/iss1/8 (aufgerufen am: 21/06/2025).

Dastin, J. (2018): Amazon Scraps Secret AI Recruiting Tool That Showed Bias Against Women, in: Reuters, 10. Oktober 2018, URL: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G (aufgerufen am: 14/06/2025).

Dörr, S. (2021): KI verlangt Corporate Digital Responsibility (CDR), in: Knappertsbusch, I. / Gondlach, K. (Hrsg.): Arbeitswelt und KI 2030, Wiesbaden: Springer Gabler, DOI: 10.1007/978-3-658-35779-5_5 (aufgerufen am: 21/06/2025).

– (2025): Praxisleitfaden Corporate Digital Responsibility: Unternehmerische Verantwortung und Nachhaltigkeitsmanagement im Digitalzeitalter, Berlin: Springer Gabler, DOI: 10.1007/978-3-662-69650-7 (aufgerufen am: 14/06/2025).

Dörr, S. / Frick, T. / Joynson, C. / Price, R. / Wade, M. (2021): The International CDR Manifesto, URL: https://corporatedigitalresponsibility.net/cdr-manifesto (aufgerufen am: 14/06/2025).

Dwork, C. / Hardt, M. / Pitassi, T. / Reingold, O. / Zemel, R. (2012): Fairness Through Awareness, in: Proceedings of Innovations in Theoretical Computer Science (ITCS), DOI: 10.1145/2090236.2090255 (aufgerufen am: 14/06/2025).

Elliott, K. / Price, R. / Shaw, P. / Spiliotopoulos, T. / Ng, M. / Coopamootoo, K. / van Moorsel, A. (2021): Towards an Equitable Digital Society: Ethics and Fairness in Algorithmic systems, in: Society, Jg. 58 / Nr. 3, 179–188, DOI: 10.1007/s12115-021-00594-8 (aufgerufen am: 21/06/2025).

Feldman, M. / Friedler, S. A. / Moeller, J. / Scheidegger, C. / Venkatasubramanian, S. (2015): Certifying and Removing Disparate Impact, in: Proceedings of the 21st ACM SIGKDD

International Conference on Knowledge Discovery and Data Mining (KDD), 259–268, DOI: 10.48550/arXiv.1412.3756 (aufgerufen am: 14/06/2025).

Flink, C. / Gross, L. / Pasmore, W. (2024): Doing Well and Doing Good. Human-Centered Digital Transformation Leadership, Singapore: World Scientific Publishing Co.

Gebru, T. / Morgenstern, J. / Vecchione, B. / Vaughan, J. W. / Wallach, H. / Daumé, H. III / Crawford, K. (2021): Datasheets for Datasets, in: Communications of the ACM, Jg. 64 / Nr. 12, 86–92, DOI: 10.1145/3458723 (aufgerufen am: 14/06/2025).

Hardt, M. / Price, E. / Srebro, N. (2016): Equality of Opportunity in Supervised Learning, in: Lee, D. / Sugiyama, M. / Luxburg, U. / Gyon, I. / Garnett, R. (Hrsg.): Advances in Neural Information Processing Systems (NeurIPS 2016), DOI: 10.48550/arXiv.1610.02413 (aufgerufen am: 14/06/2025).

Herden, C. J. / Alliu, E. / Wendt, K. / Bilgram, V. (2021): Corporate Digital Responsibility – New responsibilities in the digital age, in: Sustainability Management Forum, Jg. 29 / Nr. 1, 13–29, DOI: 10.1007/s00550-020-00509-x (aufgerufen am: 21/06/2025).

ISO / IEC (2023): ISO/IEC 42001:2023: Artificial Intelligence Management System, URL: https://www.iso.org/standard/81230.html (aufgerufen am: 21/06/2025).

Kunz, W. / Wirtz, J. (2023): Corporate Digital Responsibility (CDR) in the Age of AI – Implications for Interactive Marketing, in: Journal of Research in Interactive Marketing, Jg. 18 / Nr. 1, 31–37, DOI: 10.1108/JRIM-06-2023-0176 (aufgerufen am: 21/06/2025).

Lambrecht, A. / Tucker, C. (2019): Algorithmic Bias? An Empirical Study of Apparent Gender-Based Discrimination in the Display of STEM Career Ads, in: Management Science, Jg. 65 / Nr. 7, 2966–2981, DOI: 10.1287/mnsc.2018.3093 (aufgerufen am: 14/06/2025).

Lamont, J. / Favor, C. (2008): Distributive Justice, in: Zalta, E. N. (Hrsg.): The Stanford Encyclopedia of Philosophy, URL: https://plato.stanford.edu/entries/justice-distributive/ (aufgerufen am: 21/06/2025).

Linux Foundation (2024): AI Fairness 360, URL: https://ai-fairness-360.org/ (aufgerufen am: 21/06/2025).

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 (aufgerufen am: 21/06/2025).

Mitchell, M. / Wu, S. / Zaldivar, A. / Barnes, P. / Vasserman, L. / Hutchinson, B. / Spitzer, E. / Raji, I. D. / Gebru, T. (2019): Model Cards for Model Reporting, in: Proceedings of FAT (Fairness, Accountability, and Transparency), New York: Association for Computing Machinery, 220–229, DOI: 10.48550/arXiv.1810.03993 (aufgerufen am: 14/06/2025).

Nepomuceno, T. / Petrillo, F. (2025): The AI Fairness Myth: A Position Paper on Context-Aware Bias, in: Proceedings of the 47th International Conference on Software Engineering (ICSE), DOI: 10.48550/arXiv.2505.00965 (aufgerufen am: 21/06/2025).

Raji, I. D. / Smart, A. / White, R. N. / Mitchell, M. / Gebru, T. / Hutchinson, B. / Smith-Loud, J. / Theron, D. / Barnes, P. (2020): Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing, in: Proceedings of FAT (FAccT), DOI: 10.1145/3351095.3372873 (aufgerufen am: 14/06/2025).

Rohde, F. / Wagner, J. / Reinhard, P. / Petschow, U. / Meyer, A. / Voß, M. / Mollen, A. (2021): Nachhaltigkeitskriterien für künstliche Intelligenz. Entwicklung eines Kriterien- und Indikatorensets für die Nachhaltigkeitsbewertung von KI-Systemen entlang des Lebenszyklus, Berlin: Institut für ökologische Wirtschaftsforschung, URL: https://algorithmwatch.org/en/sustain/ (aufgerufen am: 21/06/2025).

Suresh, H. / Guttag, J. (2021): A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle, in: Proceedings of the 1st ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO '21), Article 17, New York: Association for Computing Machinery, DOI: 10.1145/3465416.3483305 (aufgerufen am: 21/06/2025).

Verma, S. / Rubin, J. S. (2018): Fairness Definitions Explained, in: 2018 IEEE/ACM International Workshop on Software Fairness (FairWare), 1–7, URL: https://fairware.cs.umass.edu/papers/Verma.pdf (aufgerufen am: 21/06/2025).