Corporate Digital Responsibility Initiative

Alexander Brink [Hrsg.]

Fairness in Zeiten Künstlicher Intelligenz




Nomos

Corporate Digital Responsibility Initiative

Alexander Brink [Hrsg.]

Fairness in Zeiten Künstlicher Intelligenz

Nomos

Die Open-Access-Veröffentlichung erfolgte mit Unterstützung durch den
Open Access Publikationsfonds der Universität Bayreuth.

Titelbilder: Erstellt durch DALL·E 2 über ChatGPT

Redaktion:
Leonhard Holzinger, Innovation Lab Ethik und Management, Universität Bayreuth, Bayreuth
Paula Becker, Innovation Lab Ethik und Management, Universität Bayreuth, Bayreuth

Eine Initiative des

Bundesministerium
der Justiz und
für Verbraucherschutz

Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.d-nb.de abrufbar.

1. Auflage 2025

Publiziert von
Nomos Verlagsgesellschaft mbH & Co. KG
Waldseestraße 3–5 | 76530 Baden-Baden
www.nomos.de

Gesamtherstellung:
Nomos Verlagsgesellschaft mbH & Co. KG
Waldseestraße 3–5 | 76530 Baden-Baden

ISBN (Print): 978-3-7560-3469-7
ISBN (ePDF): 978-3-7489-6503-9

DOI: https://doi.org/10.5771/9783748965039

Onlineversion
Nomos eLibrary

Grußwort

Liebe Leserinnen und Leser,

die digitale Transformation verändert unsere Gesellschaft grundlegend. Künstliche Intelligenz und algorithmische Systeme durchdringen immer mehr Bereiche unseres Lebens – vom Zugang zu Informationen bis hin zu Entscheidungen über Kredite oder Versicherungen. Doch während diese Technologien enorme Chancen bieten, stellen sie uns gleichzeitig vor große Herausforderungen: Wie stellen wir sicher, dass digitale Systeme fair und gerecht sind? Wie verhindern wir Diskriminierung durch Künstliche Intelligenz? Und wie können Unternehmen Verantwortung übernehmen, um eine faire digitale Zukunft zu gestalten?

Die Corporate Digital Responsibility (CDR)-Initiative des BMJV hat Fairness als zentrales Prinzip aufgenommen. Digitale Verantwortung bedeutet, nicht nur wirtschaftliche und technologische Potenziale zu nutzen, sondern auch gesellschaftliche Werte wie Gleichberechtigung, Chancengerechtigkeit und Transparenz aktiv zu fördern. Doch was bedeutet Fairness in der digitalen Welt konkret? Vier Aspekte sind aus meiner Sicht besonders entscheidend:

1. Fairness bedeutet Vielfalt und Inklusion. Digitale Anwendungen müssen die Vielfalt unserer Gesellschaft widerspiegeln und Diskriminierung aktiv vermeiden. Eine durch die CDR-Initiative in Auftrag gegebene Studie zeigt: Viele Bürgerinnen und Bürger fürchten, dass Entscheidungen von KI-Systemen soziale Ungleichheiten verstärken könnten. Um diesem Risiko zu begegnen, müssen Unternehmen sicherstellen, dass ihre Systeme auf Vielfalt getestet werden und keine verzerrten oder diskriminierenden Entscheidungen treffen.
2. Fairness erfordert Transparenz und Nachvollziehbarkeit. Viele Entscheidungen, die von KI-Systemen getroffen werden, bleiben für die betroffenen Menschen intransparent. Doch Fairness kann nur dann gewährleistet werden, wenn nachvollziehbar ist, auf welcher Grundlage Entscheidungen getroffen werden. Unternehmen sollten daher verständliche Erklärungen zu ihren digitalen Systemen liefern und Nutzerinnen und Nutzer über ihre Rechte aufklären.
3. Fairness schließt manipulative Mechanismen aus. Manipulatives Design, süchtig machende Mechanismen oder intransparente Geschäftsmodelle stehen im Widerspruch zur digitalen

Verantwortung. Unternehmen müssen darauf achten, dass ihre digitalen Angebote Nutzerinnen und Nutzern ermöglichen, souveräne und informierte Entscheidungen zu treffen, anstatt sie unbemerkt zu beeinflussen.

4. Fairness setzt gesellschaftliche Zusammenarbeit voraus. Die Entwicklung und Regulierung digitaler Technologien darf nicht in isolierten Silos erfolgen. Vielmehr müssen Politik, Wissenschaft, Wirtschaft und Zivilgesellschaft gemeinsam daran arbeiten, Standards für eine faire Digitalisierung zu etablieren. Durch den Dialog und die Zusammenarbeit verschiedener Akteure können nachhaltige Lösungen gefunden werden.

Unter dem Leitmotiv „Fairness im Zeitalter der Künstlichen Intelligenz" hat die CDR-Initiative des BMJV am 14. November 2024 auf ihrer Jahreskonferenz mit Expertinnen und Experten aus Politik, Wissenschaft, Wirtschaft und Zivilgesellschaft über diese Herausforderungen diskutiert. Im Fokus stand die Frage, wie Unternehmen ihrer digitalen Verantwortung gerecht werden und Fairness als zentrales Prinzip in ihre Strategien integrieren können. Die Konferenz beleuchtete praktische Lösungsansätze und bot eine Plattform für den Austausch über innovative Ideen und Best Practices.

Ich hoffe, dass die Diskussionen dieser Konferenz wertvolle Impulse liefern und Unternehmen ermutigen, Fairness als Leitprinzip der digitalen Transformation zu verankern. Denn nur gemeinsam können wir eine digitale Zukunft gestalten, die allen Menschen zugutekommt.

Frank Schwabe

Parlamentarischer Staatssekretär bei der Bundesministerin der Justiz und für Verbraucherschutz

Bericht zur Konferenz #CDRK24

Alexander Brink

Am 14. November 2024 fand die fünfte CDR-Konferenz der Corporate Digital Responsibility (CDR) Initiative unter dem Titel „Fairness im Zeitalter der Künstlichen Intelligenz" im Data Space in Berlin statt. Die Veranstaltung brachte Expert*innen aus Politik, Wirtschaft und Zivilgesellschaft zusammen, um die Bedeutung von Fairness in der digitalen Transformation zu erörtern und praxisnahe Lösungsansätze vorzustellen. Mit inspirierenden Vorträgen, interaktiven Workshops und ausreichend Raum für Networking bot die Konferenz den rund 150 Teilnehmenden eine Plattform für intensiven Austausch und Diskussionen.

Eröffnung und Rahmen des Themas von Prof. Dr. Christian Thorun (13.00–13.10 Uhr)

Die Veranstaltung wurde mit einer Begrüßung und einer thematischen Einführung eröffnet, gefolgt von einem Überblick über die Agenda. Prof. Dr. Christian Thorun betonte die Bedeutung der Integration von CDR-Prinzipien in die Unternehmensstrategie, insbesondere im Hinblick auf Fairness, um eine ausgewogene und gerechte digitale Transformation zu gewährleisten.

Grußwort von Dr. Christiane Rohleder (13.10–13.25 Uhr)

Es folgte ein Grußwort von Staatssekretärin Dr. Christiane Rohleder vom Bundesministerium für Umwelt, Naturschutz, nukleare Sicherheit und Verbraucherschutz (BMUV). Sie betonte die zentrale Rolle von Fairness in der digitalen Transformation und die Verantwortung von Unternehmen, gerechte und transparente KI-Systeme zu entwickeln und einzusetzen.

Keynote von Dr. Till Klein (13:25–13:45 Uhr)

Dr. Till Klein, Leiter der AI-Regulierung beim appliedAI Institute for Europe, hielt die Eröffnungsrede zum Thema „Fairness entlang der KI-Wertschöpfungskette: Was ändert sich mit dem AI Act?". Er erläuterte die bevorstehenden regulatorischen Änderungen durch den AI Act und

deren Auswirkungen auf Unternehmen. Dabei hob er hervor, wie wichtig es ist, Fairness in allen Phasen der KI-Entwicklung und -Anwendung zu berücksichtigen.

Paneldiskussion: Herausforderungen und Lösungsansätze für Fairness in der Digitalisierung (13:45–14:30 Uhr)

In der anschließenden Paneldiskussion diskutierten Dr. Christiane Rohleder (BMUV), Johanna Hinz (Leiterin der Tech Strategy bei der Otto Group), Elisabeth Nöfer (Projektmanagerin Digitalisierte Gesellschaft bei der Stiftung Mercator) und Dr. Till Klein über die praktischen Herausforderungen bei der Umsetzung von Fairness in digitalen Prozessen. Themen wie die Definition von Fairness, die Umsetzung regulatorischer Anforderungen und der Umgang mit möglichen Zielkonflikten standen im Mittelpunkt der Diskussion.

Vorstellung der Diversity-Folgenabschätzung (14:30–14:40 Uhr)

Im Anschluss präsentierten Maria Hinz (BARMER) und Thomas Bieler (ING Deutschland) die neu entwickelte Diversity-Folgenabschätzung. Dieses Tool unterstützt Unternehmen dabei, die Auswirkungen von Digitalisierungsprozessen auf Diversität, Inklusion und Chancengleichheit zu evaluieren und entsprechende Maßnahmen zur Förderung dieser Werte zu identifizieren.

Ausblick auf die Workshops (14.40–14.50 Uhr)

Im Gespräch zwischen Prof. Dr. Christian Thorun und Prof. Dr. Dr. Alexander Brink wurde der Band „Menschenzentrierte Digitalisierung" der CDR-Initiative vorgestellt, an dem sich ca. 30 Expert*innen beteiligten. Es folgte ein kurzer Überblick über die anstehenden Workshops zu spezifischen Fairness-Aspekten in verschiedenen Bereichen der digitalen Technologie.

Kaffeepause (14:50–15:20 Uhr)

Die Kaffeepause bot den Teilnehmenden die Möglichkeit, sich untereinander auszutauschen, Netzwerke zu knüpfen und die bisherigen Eindrücke zu reflektieren. Der Data Space Berlin erwies sich dabei als idealer Ort, der mit seiner modernen und inspirierenden Atmosphäre den offenen Dialog und kreative Diskussionen förderte.

Workshops: Praxisnahe Einblicke in CDR-Themen (15:20–16:15 Uhr)

In vier parallelen Workshops wurden spezifische Themen vertieft:

- Workshop 1: „Diversity-Folgenabschätzung: Ein Instrument zur Förderung von Diversität, Inklusion und Chancengleichheit" – mit Franziska Nitsche (Otto Group), Thomas Bieler (ING Deutschland) und Prof. Dr. Christian Thorun (CDR-Initiative).
- Workshop 2: „Fairness in KI-Richtlinien: Durchsetzbar oder leeres Versprechen?" – mit Sebastian Reimann (XING – Teil von NEW WORK SE) und Leonhard Henke (CDR-Initiative).
- Workshop 3: „Künstliche Intelligenz im Einsatz: Chancen und Herausforderungen von internen KI-Assistenten mit GPT-Modellen in der Praxis" – mit Anja Koerber (Otto Group) und Dr. Sara Elisa Kettner (CDR-Initiative).
- Workshop 4: „Desinformation und Hate Speech: Welche Rolle haben Unternehmen bei der Gestaltung eines fairen gesellschaftlichen Diskurses?" – mit Maria Hinz (BARMER), Deniz Taskiran (Telefónica) und Dr. Frank Esselmann (CDR-Initiative).

Blitzlichter aus den Workshops (16:15–16:30 Uhr)

Nach den Workshops wurden die wichtigsten Erkenntnisse und Diskussionsergebnisse in kurzen Präsentationen zusammengefasst, um allen Teilnehmenden im Plenum einen umfassenden Überblick über die behandelten Themen zu geben.

Abschluss und Ausblick (16:30–17:00 Uhr)

Zum Abschluss der Konferenz berichteten Benjamin Hoffmann (Atruvia) und Dr. Stefan König (Markgraf) über ihre Erfahrungen als neue Mitglieder der CDR-Initiative. Sie hoben die Vorteile des interdisziplinären Austauschs und der Zusammenarbeit innerhalb der Initiative hervor. Dr. Frank Esselmann (CDR-Initiative) gab einen Ausblick auf kommende Aktivitäten und lud interessierte Unternehmen ein, sich der Initiative anzuschließen, um gemeinsam Verantwortung in der digitalen Transformation zu übernehmen.

Die CDR-Konferenz 2024 bot eine wertvolle Plattform für den Austausch über Fairness in der digitalen Welt und zeigte praxisnahe Wege auf, wie Unternehmen ihrer digitalen Verantwortung gerecht werden können.

Inhaltsverzeichnis

1. GRUNDLAGEN

2. VERBRAUCHERPERSPEKTIVE

3. GOVERNANCE

4. IMPULSE

1.

Grundlagen

Von der Ethik zum Gesetz[1]

Wie der EU AI Act die Fairness entlang der KI-Wertschöpfungskette formalisiert

Till Klein

1. Einführung

Die Künstliche Intelligenz (KI) durchdringt zunehmend alle Bereiche unseres gesellschaftlichen und wirtschaftlichen Lebens. Von der automatisierten Kreditvergabe in Banken über die Vergabe von Kindergartenplätzen bis hin zur Personalauswahl in Unternehmen – KI-Systeme unterstützen oder treffen Entscheidungen, die das Leben von Menschen maßgeblich beeinflussen. Mit der wachsenden Bedeutung von KI rückt auch die Frage nach der Fairness dieser Systeme immer stärker in den Fokus von Politik, Wirtschaft und Gesellschaft.

Die Europäische Union hat mit der KI-Verordnung (engl. „AI Act") einen wegweisenden regulatorischen Rahmen geschaffen, der erstmals verbindliche Regeln für den Umgang mit Künstlicher Intelligenz etabliert. Der AI Act trat am 1. August 2024 in Kraft und schützt neben der Gesundheit und Sicherheit auch die Grundrechte der Bürgerinnen und Bürger, inklusive dem Schutz vor Diskriminierung. Bei der Umsetzung stellen sich konkrete Fragen:

- Welche Anforderungen sieht der AI Act bezüglich Fairness vor?
- Wie verteilen sich diese Anforderungen auf die KI-Wertschöpfungskette?
- Welche Herausforderungen stellen sich dabei?

---

1 KI wurde für einen initialen Entwurf des Textes auf Grundlage der Folien der Keynote von 2024 verwendet, dann aber verworfen und komplett neu geschrieben. Ein paar wenige Sätze in der Einleitung sind nah am Text der KI.

Dieser Beitrag untersucht, wie sich die neuen rechtlichen Anforderungen im AI Act auf die verschiedenen Stufen der KI-Entwicklung und -Anwendung auswirken – von den Anbietern von Basismodellen über die Entwickler von KI-Systemen bis hin zu den Betreibern und letztendlich den betroffenen Personen.[2] Dabei wird deutlich, dass Fairness in KI-Systemen kein isoliertes technisches Problem darstellt, sondern eine umfassende Aufgabe, die Anstrengungen und Zusammenarbeit der beteiligten Akteurinnen und Akteuren entlang der KI-Wertschöpfungskette erfordert.

2. Fairness als Feature

2.1 Wie KI-Systeme diskriminieren

Fairness in KI-Systemen zu definieren ist komplex und vielschichtig, insbesondere mit Blick auf die interdisziplinäre Natur der Sache. Mathematiker haben ein unterschiedliches Verständnis von Bias im Vergleich zu Soziologen oder Juristen, jedoch sind diese und weitere Rollen häufig bei der Entwicklung und Nutzung von KI-Systemen involviert.

Nach der IEC Norm zu „Bias in KI-Systemen" (ISO/IEC TR 24027:2021) bedeutet Fairness, unparteiisch zu sein und sich ohne Bevorzugung oder Diskriminierung zu verhalten (vgl. ISO/IEC 2021). Der Gegenpol von Fairness ist die Voreingenommenheit (engl. „bias"). Voreingenommenheit beschreibt die systematisch unterschiedliche Behandlung bestimmter Objekte, Personen oder Gruppen im Vergleich zu anderen. Eine ungleiche Behandlung ist per se nicht „schlecht", denn in vielen praktischen Anwendungen werden KI-Systeme explizit entwickelt, um Dinge oder Personen unterschiedlich zu behandeln, etwa in der Qualitätskontrolle, beim Sortieren (z. B. Güteklassen von Obst und Gemüse) oder in der Diagnostik (z. B. bei Krankheitsbildern, die sich in Abhängigkeit von Geschlecht oder Alter unterschiedlich zeigen). Diskriminierung durch KI findet statt, wenn die Voreingenommenheit in einem KI-System zu einer ungerechten oder vorurteilsbehafteten Behandlung von Einzelpersonen, Organisationen, Gruppen oder Gesellschaften führt. Es gibt zahlreiche Beispiele in denen Personen aufgrund ihrer Herkunft, Sprache, Religion oder sexuellen Orientierung benachteiligt behandelt werden, und KI-Systeme bergen das Potential diese Diskriminierung aufzunehmen, zu

[2] Der AI Act spricht von ‚KI-Modellen mit allgemeinem Verwendungszweck' (engl. „General Purpose AI Models", kurz: GPAI-Model).

verstärken und zu skalieren, insbesondere wenn die Trainingsdaten diese Voreingenommenheit widerspiegeln. Allgemein sind die Ursachen von Voreingenommenheit in KI-Systemen ist vielfältig und liegt insbesondere in den frühen Phasen der Systementwicklung (vgl. Norori et al. 2021):

- „Datenbedingter Bias" entsteht durch Trainingsdaten, die nicht repräsentativ für das Phänomen sind, das sie beschreiben sollen, etwa durch überproportional häufiges Vorkommen bestimmter Proben oder Lücken. In diesem Fall lernt das KI-Model diese Verzerrungen und reproduziert sie in seinen Entscheidungen.
- „Algorithmischer Bias" kann durch unausgewogene Klassen oder systematische Fehler im Trainingsprozess entstehen. Auch die Wahl bestimmter Algorithmen, Parameter oder Optimierungsziele kann zu diskriminierenden Ergebnissen führen.
- „Menschlicher Bias" fließt durch gesellschaftliche Vorurteile und Machtungleichgewichte in die Systemgestaltung ein, zum Beispiel durch die Ansichten von Entwicklern, Datenannotatierern[3] und Entscheidungsträgerinnen, die bewusst oder unbewusst ihre eigenen Voreinstellungen mit einbringen.

Das Ergebnis ist ein verzerrter Output, der bestimmte Gruppen systematisch benachteiligt oder bevorzugt. Diese Verzerrungen können sich besonders problematisch auswirken, wenn KI-Systeme in sensiblen Bereichen wie der Kreditvergabe, Personalauswahl oder Strafverfolgung eingesetzt werden.

2.2 Ob KI zu mehr Fairness oder Unfairness führt, ist eine Frage der Umsetzung

Neben der technischen Umsetzung eines KI-Systems hängt es vor allem von der praktischen Umsetzung und dem jeweiligen Kontext ab, ob KI zu mehr oder weniger Fairness führt, was die folgenden Beispiele illustrieren. Ein positives Beispiel ist das KI-System KitaMatch aus Deutschland. Das System adressiert eine gängige Herausforderung junger Eltern bei der Suche nach einem Kita-Platz für den Nachwuchs. Um die eigenen Chancen auf einen Platz zu erhöhen, stellt man Anträge nicht bei einer, sondern bei mehreren Kitas. Die Kitas wiederum beobachten durch die Mehrfachanmeldung eine aufgeblasene Nachfrage, was zu langen Wartelisten, vielen Absagen und

[3] Annotieren bedeutet, Daten mit Anmerkungen oder Notizen zu versehen. Im Bereich KI spricht man umgangssprachlich von „Labeln", also der Markierung von Trainingsdaten durch Menschen, von denen das KI-System lernt.

einem erhöhten Koordinationsaufwand führt. Das KI-System KitaMatch soll hier durch eine Matching-Anwendung Abhilfe schaffen.

> KitaMatch ist ein Verfahren zur fairen, schnellen und transparenten Vergabe von Kitaplätzen, das Eltern und Kitas bei der Vergabe von Betreuungsplätzen optimal miteinander verbindet (KitaMatch 2023).

Die Kitas profitieren unter anderem von „gerichtsfesten" Gründen für die Vergabe, inklusive klare Argumente (gemäß einem Kriterienkatalog), warum andere Kinder priorisiert wurden. Die Jugendämter können sich sicher sein, dass Kinder einheitlich und nachvollziehbar priorisiert wurden. Die Eltern erhalten eine transparente Erklärung, falls das eigene Kind keinen Platz erhält und die standardisierte Vorgehensweise beschleunigt den gesamten Prozess, sodass Eltern sich einen Plan B zurechtlegen können.

Im Kontrast dazu steht das französische KI-System namens „Parcoursup", das zur Vergabe von Universitätsplätzen verwendet wird und dabei zu struktureller Diskriminierung geführt hat (vgl. Federal Anti-Discrimination 2019). Das System gleicht die Bewerbungen von Schulabsolventinnen und Schulabsolventen mit verfügbaren Plätzen an Universitäten auf nationaler Ebene ab, während lokale Universitäten zusätzlich eigene Auswahl-Algorithmen verwenden. Schulabsolventen versenden im Durchschnitt 10–20 Bewerbungen an Universitäten im ganzen Land, wobei sie persönliche Daten angeben müssen, darunter den Wohnort, das Einkommen und die vorherige Schule. Problematisch wurde hier die mangelnde Transparenz: der nationale Algorithmus war öffentlich zugänglich, jedoch nicht lokalen Sortieralgorithmen. Die Angabe von Einkommen und Wohnort führte zu einer systematischen Benachteiligung weniger wohlhabenden Bewerber und Bewerberinnen oder solche aus bestimmten (Vor-)Orten.

Die französische Gleichstellungsbehörde Défenseur des Droits führte eine Untersuchung durch und kritisierte besonders die Verwendung der besuchten Schule als Auswahlkriterium, da dies aufgrund geografischer Lage zu einer Diskriminierung führen kann. Dieser Fall verdeutlicht die Komplexität entlang der KI-Wertschöpfungskette, denn es bedarf ein Zusammenspiel unterschiedlicher Akteurinnen und Akteure und deren Systeme, um die transparente und verantwortungsvolle Nutzung von KI zu gewährleisten. Dieses Beispiel demonstriert außerdem, wie scheinbar neutrale technische Systeme gesellschaftliche Ungleichheiten verstärken können, wenn Transparenz und systematische Fairness-Prüfungen fehlen.

2.3 Bias in „großen Modellen" multipliziert sich entlang der KI-Wertschöpfungskette

Die Popularität „großer Modelle", die meist von großen Tech-Firmen wie OpenAI, META, Amazon oder Google zur Verfügung gestellt werden, stellen eine besondere Herausforderung für Transparenz und Fairness dar, weil sie für den Mangel an Transparenz kritisiert werden und gleichzeitig bei zigtausenden nachgelagerten Anbietern (engl. „Downstream-Provider") zum Einsatz kommen. Ein plakatives Beispiel für Bias in Basismodellen ist die Studie „Which Humans?" von Atari et al. aus dem Jahr 2023. Die Untersuchung beginnt mit der Prämisse, dass die Anbieterinnen und Anbieter dieser Modelle gerne von der Demokratisierung von KI sprechen und erklären, dass die neuen KI-basierten Fähigkeiten für „alle Menschen" sind. Die Autoren des Papers fragen „which humans?" („welche Menschen") denn die Ergebnisse zeigen, dass die getesteten Basismodellen einen kulturellen Bias innehaben und die Werte und Konzepte von Menschen in Abhängigkeit ihrer Herkunft mehr oder weniger gut widerspiegeln. Konkret: je „näher" die Kultur eines Landes an der Kultur der USA liegt, desto besser resonieren die Ergebnisse der Basismodelle mit den Nutzenden. Zum Beispiel sehen sich Nutzenden aus Kanada und Australien sehr gut in den Ergebnissen, während das Gegenteil für Personen aus Ägypten, Jordanien oder Pakistan gilt (vgl. Abbildung 1).

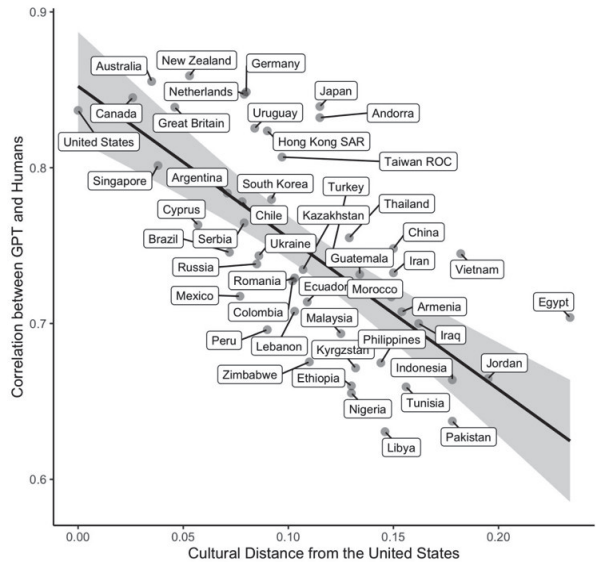


ABBILDUNG 1: KULTURELLER BIAS (QUELLE: ATARI ET AL. 2023: 11)

Sofern Bias in einem KI-Modell bekannt und transparent kommuniziert ist, können nachgelagerte Akteurinnen und Akteure nach einem geeigneten Umgang suchen und zum Beispiel ein anderes Modell verwenden oder Mitigationsmaßnahmen umsetzen. Hier liegt eine Herausforderung, denn erstens ist es häufig unklar ob bzw. welche Schwächen an Basismodellen bekannt sind (selbst bei den eigenen Anbietern, siehe „Emerging Abilities" [vgl. Berti et al. 2025]) und zweitens werden die Model-Anbieter häufig für einen Mangel an Transparenz kritisiert. Der AI Safety Report bewertet führende Basismodelle anhand bestimmter Kriterien und kommt zu folgendem Fazit (übersetzt aus dem Englischen): Trotz des wachsenden internationalen Konsenses über die Risiken der KI und der zunehmenden Belege für rasante Fortschritte bei den Fähigkeiten warnen Experten, dass die Kluft zwischen technologischen Ambitionen und Sicherheitsvorkehrungen immer größer wird. Unternehmen streben nach künstlicher allgemeiner Intelligenz und sagen voraus, dass sie innerhalb dieses Jahrzehnts übermenschliche Leistungen erzielen werden. Doch wie ein Prüfer feststellte, „hat keines der Unternehmen einen kohärenten, umsetzbaren Plan" zur Kontrolle solcher Systeme (vgl. Future of Life Institute 2025: 21).

Ungeachtet der oben genannten Mängel hat die Nutzung von generativer KI seit ChatGPT rasant zugenommen, weil wesentliche Adaptions-barrieren gesunken sind. Die Benutzerschnittstelle ist einfach zu bedienen, Ein- und Ausgabe erfolgen in natürlicher Sprache, und Rechenleistung und Speicher werden über die Plattform im Hintergrund abgedeckt. Laut einer Studie des Start-Up Verbands (vgl. Startup Verband 2024) haben 43% der befragten Unternehmen GenAI regelmäßig im Einsatz und weitere 35% nutzen es im operativen Alltag. Mit jedem Einsatz, jeder Adaption, werden die Verzerrungen und Mängel der generativen Modelle weiter multipliziert und verbreitet.

2.4 Fairness in KI als Weg zu mehr Wettbewerbsfähigkeit?

„Ohne Vertrauen, keine Nutzung" (engl. „no trust, no use") – so lautet eine gängige Formel der Befürworterinnen und Befürworter von vertrauenswürdiger KI. Wenn wir einem KI-System nicht trauen können, zum Beispiel für die Korrektheit oder Nachvollziehbarkeit der Ergebnisse, dann werden wir es kaum (für kritische Anwendungen) einsetzen. Der Umkehrschluss suggeriert, dass Vertrauenswürdigkeit ein wesentlicher Faktor ist, um die Adaption von KI zu erhöhen, also: „mehr Vertrauen, mehr Nutzung".

Eine Umfrage des Digital-Verbands Bitkom (2024) zeigt, dass ein überschaubarer Anteil von 9% der befragten 602 Unternehmen GenAI bereits nutzen (vgl. Bitkom 2024). Weitere 18% planen den Einsatz und weitere 19% können es sich vorstellen. Demgegenüber stehen die privaten Nutzer und Nutzerinnen, bei denen bereits 30% der deutschen Gesamtbevölkerung allein ChatGPT aktiv nutzen, wie eine Umfrage des Nürnberg Institut für Marktentscheidungen e.V. berichtet (vgl. Kaiser et al. 2024). Insbesondere bei etablierten Unternehmen gibt es Zurückhaltung bei Investitionen in die Nutzung von GPAI-Modellen, unter anderem wegen Bedenken bei ihrer Zuverlässigkeit. Diese Daten deuten auf einen Mangel an Vertrauen gegenüber generativer KI in etablierten Unternehmen.

Zahlreiche Umfragen setzen hier an und beleuchten die Hindernisse für die Nutzung von KI bzw. generativer KI, wobei die Leistungsfähigkeit der Modelle ein wiederkehrender Aspekt ist. Unter den Top 5 Gründen der Bitkom-Umfrage nennen 65% der Unternehmen „Schlechte Qualität der Ergebnisse" als Hemmnis für den Einsatz von generativer KI (vgl. Bitkom 2024). Eine Umfrage der Expertenkommission Forschung und Innovation zeigt, dass „Bedenken hinsichtlich Reife und Zuverlässigkeit von KI" ein wesentliches Hindernis für den Einsatz von KI in Unternehmen ist (vgl. EFI 2024). Auch bei der Nutzung im privaten Umfeld unterstreicht das Eurobarometer zur Digitalen Dekade der EU, dass sich die Mehrheit der Befragten digitale Dienstleistungen wünschen, die besser auf die persönlichen Bedürfnisse abgestimmt sind.

In dem Korridor zwischen „Bias in GPAI Modellen" und „Keine Nutzung von GenAI wegen schlechter Qualität der Ergebnisse" liegt die Chance robuste Modelle zu bauen, die alle Nutzergruppen „gleich gut" behandeln, unabhängig von ihrem Herkunftsland, die zuverlässige Ergebnisse erzeugen und die den Mensch in den Mittelpunkt stellen. Wenn dies gelingt, können Fairness, und andere Aspekte vertrauenswürdiger KI, ein Erfolgsfaktor für die breitere Nutzung von KI in Unternehmen und im persönlichen Umfeld sein. Die Beispiele von KitaMatch und Parcoursup aus Frankreich zeigen, dass dafür alle Akteurinnen und Akteure in der KI-Wertschöpfungskette eine Rolle spielen, vom GPAI-Anbieter, über die nachgelagerten Anbieter, bis hin zum Endnutzer. Der AI Act hat diese unterschiedlichen Rollen aufgegriffen und sieht konkrete Pflichten und Rechte für sie vor.

3. Von der Ethik zum Gesetz: Fairness in der europäischen KI-Verordnung

3.1 Ethische Richtlinien als Säule des AI Acts

Der AI Act wird mitunter als formalisierte Ethik bezeichnet, weil die neuen Regularien auf ethischen Prinzipien aufbauen und konkrete Anforderungen für die ethische Entwicklung und Nutzung von KI beinhalten. Diese Entwicklung ist konsistent mit dem Vorgehen der EU, denn bereits 2019 hat eine von der EU beauftragte Expertengruppe den Begriff „vertrauenswürdige KI" als Leitbild formuliert (vgl. EU Kommission 2019). Vertrauenswürdige KI hat drei wesentliche Merkmale (übersetzt aus dem Englischen):

1. Rechtmäßig – Einhaltung aller geltenden Gesetze und Vorschriften
2. Ethisch – Einhaltung ethischer Grundsätze und Werte
3. Robust – aus technischer Sicht und unter Berücksichtigung des sozialen Umfelds

3.2 Schutz von Grundrechten als explizites Ziel

Beim Blick in den AI Act finden sich klare Hinweise für eine ethische und Mensch-zentrierte Ausgestaltung des Regelwerkes, beginnen bei Artikel 1, dem Gegenstand:

> (1) Zweck dieser Verordnung ist es, [...] die Einführung einer auf den Menschen ausgerichteten und vertrauenswürdigen künstlichen Intelligenz (KI) zu fördern und gleichzeitig ein hohes Schutzniveau in Bezug auf Gesundheit, Sicherheit und die in der Charta verankerten Grundrechte, [...] vor schädlichen Auswirkungen von KI-Systemen in der Union zu gewährleisten und die Innovation zu unterstützen.

Die KI-Verordnung definiert damit den Schutz der Grundrechte als eine zentrale Säule und verfolgt dabei einen menschenzentrierten Ansatz für vertrauenswürdige KI. Zusätzlich zielt der Gesetzgeber darauf ab, ein hohes Schutzniveau für andere Aspekte wie Gesundheit, Sicherheit, Demokratie und Rechtsstaatlichkeit zu gewährleisten. Neben den Schutzzielen soll mit dem AI Act Innovation und wirtschaftliche Entwicklung gefördert werden, d.h. es geht nicht um Regulierung ODER Innovation, sondern um die balancierte Symbiose der vermeintlich im Konflikt stehenden

Ziele, wobei die die Grundwerte der Union, im Sinne der Grundrechte der Europäischen Union („Charta"), als normative Leitplanken fungieren. Die Charta ist Teil der Europäischen Verträge und gibt den Bürgerinnen und Bürgern der EU-Rechte, die unter anderem auf Fairness und Gleichbehandlung abzielen (vgl. Abbildung 2) (vgl. Europäische Union 2000):

Artikel Nr.	Titel
21	Nichtdiskriminierung
23	Gleichheit von Männern und Frauen
24	Rechte des Kindes
25	Rechte älterer Menschen
26	Integration von Menschen mit Behinderung

ABBILDUNG 2: BEISPIELHAFTE AUSWAHL VON GRUNDRECHTEN DER EU ZUM THEMA FAIRNESS UND GLEICHBEHANDLUNG (QUELLE: EUROPÄISCHEN UNION 2000: 13 F.)

3.3 Ethische Richtlinien als Säule des AI Acts

Die erfolgreiche Überführung ethischer Prinzipien zum Gesetz attestieren auch wissenschaftliche Aufsätze, etwa von Nathalie Smuha von der KU Leuven in Belgien (vgl. Smuha 2024). Ihre Analyse zeigt auf, dass die Empfehlungen der sogenannten High-Level Expert Group (vgl. EU Kommission 2019) zu vertrauenswürdiger KI ihren Weg in den Text des AI Acts gefunden haben. Dazu zählen vier zentrale Grundsätze, inklusive Fairness (die anderen drei Grundsätze sind: Achtung der menschlichen Autonomie, Schadensverhütung und Erklärbarkeit). Konkret schlagen sich die ethischen Richtlinien wieder in den Anforderungen an Hochrisiko-KI-Systeme (Artikel 8–15, z. B. im Bereich Data Governance) und bei der Auswahl verbotener KI-Praktiken (Artikel 5). In der Summe dienten die ethischen Richtlinien als normativer Kompass für den AI Act und stellten die Grundlage für den freiwilligen Verhaltenskodex.

Diese nachfolgende Sektion behandelt die Frage, welche Anforderungen an Fairness entlang der KI-Wertschöpfungskette im AI Act zu finden sind. Für diesen Zweck betrachten wir ein fiktives Szenario in der Finanzbranche und untersuchen anhand dessen, welche möglichen Änderungen zu erwarten sind.

4. *Fairness entlang der KI-Wertschöpfungskette*

4.1 Ein exemplarisches Szenario: Kreditwürdigkeitsprüfung (Credit Scoring)

Kreditanträge können in den unterschiedlichsten Lebenslagen notwendig sein: das erste Auto, eine Immobilie, eine Unternehmensgründung oder der große Wunsch, der noch offen war. Ein wesentlicher Schritt auf der Seite des Kreditinstitutes ist die Bonitätsprüfung, also die Prüfung, ob und zu welchen Bedingungen ein Kredit vergeben werden kann. Im Hintergrund steht die Frage der Ausfallwahrscheinlichkeit: Wird der Kreditnehmer den Betrag vollständig und fristgerecht zurückzahlen können? Um diese Prüfung zu automatisieren und zu beschleunigen, kommen immer häufiger KI-Systeme zum Einsatz, die anhand bestimmter Eingabewerte, die Ausfallwahrscheinlichkeit des Antragstellers bewerten und damit die Entscheidung zur Kreditvergabe wesentlich beeinflussen können. Stellen wir uns vor, ein junges Paar stellt einen Kreditantrag (z. B. für den Kauf einer Immobilie) und die Finanzberaterin erklärt, dass die Kreditwürdigkeit mithilfe eines KI-Systems ermittelt wurde. Bei der Besprechung der Konditionen des Kredits, die überraschend ungünstig ausfallen, beschleicht das Paar ein seltsames Gefühl und es fragt nach, wie der Bonitätswert zustande kam.

Mit Blick auf den AI Act, ist das Kreditinstitut in diesem Beispiel der Betreiber des KI-Systems und das Paar sind die betroffenen Personen. Der Anbieter des KI-Systems ist möglicherweise eine andere Entität, z. B. ein FinTech Startup, das unter anderem ein großes Sprachmodell[4] nutzt, mit den Informationen aus dem Internet und den Sozialen Medien über die Antragsteller ausgewertet werden, etwa mit Blick auf „gefährliche Hobbies" oder andere Verhaltensweisen, die die Kreditwürdigkeit beeinflussen könnten. Basierend auf der KI-Wertschöpfungskette, bestehend aus GPAI-Model Anbieter, KI-System Anbieter, KI-System Betreiber und betroffener Personen, stellt sich die Frage: Welche Pflichten oder Rechte des AI Acts tragen zu mehr Fairness für die betroffenen Personen bei?

Die folgenden Absätze betrachten verschiedene Rollen entlang der KI-Wertschöpfungskette und (a) zeigen zentrale Anforderungen aus dem AI Act zu Fairness auf und (b) weisen auf offene Punkte hin (Stand Mitte 2025).

---

[4] Nehmen wir an, beim Sprachmodell handelt es sich um ein „KI-Modell mit allgemeinem Verwendungszweck" (GPAI-Model) im Sinne des AI Acts.

Die betroffene Person ist die Person, die von dem Ergebnis oder einer Entscheidung eines KI-Systems betroffen ist. Im oben genannten Beispiel ist es das Paar, denn die Bedingungen für ihren individuellen Kredit werden von dem KI-System für die Kreditwürdigkeitsprüfung beeinflusst. Der AI Act enthält Rechtsbehelfe (Kapitel 9, Abschnitt 4), die die Position von betroffenen Personen stärken. Konkret geht es um das Recht auf Erläuterung (Artikel 86), wonach betroffene Personen von Hochrisiko-KI-Systemen im Fall eines begründeten Verdachts das Recht haben „vom Betreiber eine klare und aussagekräftige Erläuterung zur Rolle des KI-Systems im Entscheidungsprozess und zu den wichtigsten Elementen der getroffenen Entscheidung zu erhalten". Falls Grund zur Annahme besteht, dass gegen Bestimmungen des AI Acts verstoßen wurde, können gemäß Artikel 85 natürliche und juristische Personen „bei der betreffenden Marktüberwachungsbehörde Beschwerden einreichen". Im Sinne des Beispiels könnte das Paar vom Kreditinstitut eine Erläuterung verlangen, da das KI-System zur Kreditwürdigkeitsprüfung gemäß Artikel 6 in Kombination mit Anhang III wahrscheinlich ein Hochrisiko-KI-System ist.

Die durch den AI Act eingeführten Rechte für betroffene Personen verringern das Risiko, mit „unfairen" KI-Systemen konfrontiert zu werden, weil Anbieter mit entsprechenden Rückfragen rechnen müssen. Gleichzeitig können sich betroffene Personen auf ihr Recht berufen, bzw. im Schadensfall eine Wiedergutmachung verlangen. Inwiefern Fairness tatsächlich gestärkt wird, hängt jedoch von der Umsetzung bzw. Durchsetzung dieser Rechte ab, da es in der Praxis noch offene Fragen gibt. Hier eine Auswahl:

- Begriffsdefinition: Der Begriff „betroffene Person" ist nicht im AI Act definiert, wodurch ein Interpretationsspielraum entsteht, der im Einzelfall unterschiedlich ausgelegt werden kann. Zum Beispiel ist unklar, welches Tatbestandsmerkmal erfüllt sein muss, damit eine Person als „betroffen" gilt.
- Bewusstsein: Nur wer die eigenen Rechte kennt, kann sie geltend machen. Jedoch sind die neuen Rechte durch den AI Act weitgehend unbekannt unter den potenziell betroffenen Personen (insbesondere Verbraucherinnen und Verbraucher). Bei der Datenschutz Grundverordnung (DSGVO) hat es Jahre gedauert, ein gewisses Bewusstsein in der Gesellschaft aufzubauen (vgl. Rughinis et al. 2019).
- Durchsetzung: Jeder Mitgliedstaat muss gemäß Artikel 70 bis zum 02. August 2025 eine Marktaufsichtsbehörde benennen und öffentlich bekannt geben, wie diese Behörde über

elektronische Wege erreicht werden kann. Jedoch ist diese Behörde Mitte August 2025 für Deutschland noch nicht offiziell benannt worden, das heißt, betroffene Personen können ihr „Recht auf Beschwerde“ noch nicht ausüben.

4.3 Fairness und Betreiber von KI-Systemen

Betreiber von KI-Systemen sind natürliche oder juristische Personen, die ein KI-System in eigener Verantwortung verwenden, siehe Artikel 3 (4) AI Act. Im Beispiel der Kreditwürdigkeitsprüfung ist das Kreditinstitut der Betreiber, weil die Finanzberaterin das KI-System in ihrer Rolle als Angestellte und im Auftrag des Arbeitgebers einsetzt.

Der AI Act enthält mehrere Anforderungen an Betreiber, die zu mehr Fairness für die betroffenen Personen führen sollten. Erstens sind Betreiber von bestimmten Hochrisiko-KI-Systemen, darunter auch solche für Kreditwürdigkeitsprüfungen, angehalten, eine Grundrechtefolgenabschätzung vor der Inbetriebnahme durchzuführen, siehe Artikel 27. Die zentrale Frage ist, welche Auswirkungen die Verwendung des KI-Systems auf die Grundrechte haben kann, insbesondere mit Blick auf Schadensrisiken für betroffene Personen. Zweitens definiert Artikel 26 spezifische Pflichten für Betreiber von Hochrisiko-KI-Systemen, zum Beispiel die Nutzung des KI-Systems gemäß der Betriebsanleitung des Anbieters, die Benennung einer menschlichen Aufsicht über das KI-System, sowie die Pflicht, Vorfälle mit dem KI-System an den Anbieter und die zuständigen Behörden zu melden. Außerdem müssen Arbeitnehmervertreter vor der Inbetriebnahme des KI-Systems darüber informiert werden.

Bei der Einhaltung dieser Pflichten hat das Paar im Beispiel der Kreditwürdigkeitsprüfung eine höhere Chance auf eine faire Behandlung, weil mögliche Grundrechtsverletzungen vor der Inbetriebnahme ermittelt und mitigiert wurden. Auch die Finanzberaterin sollte mit dem KI-System vertraut sein und unregelmäßige Ergebnisse erkennen bzw. korrigieren können. Dennoch gibt es auch hier offene Punkte in der praktischen Umsetzung:

- Methodik: Zur effektiven Durchführung einer Grundrechtefolgenabschätzung von KI-Systemen bedarf es allgemein anerkannter Methoden, jedoch gibt weder der AI Act diese Methode vor, noch enthält er einen Auftrag an das AI Office oder anderen Akteurinnen und Akteure, diese Methode zu entwickeln.[5]

[5] Das dänische Institut für Menschenrechte bietet eine umfassende Methode für Grundrechtefolgenabschätzungen im Digitalen Kontext (vgl. Danish Institute for Human Rights 2025).

- Vorfälle: Im Gegensatz zu Vorfällen bezüglich Gesundheit und Sicherheit sind Vorfälle an den Grundrechten schwieriger zu erkennen, da diese mitunter subtil, implizit oder akkumulativ auftreten, zum Beispiel eine täglich wiederkehrende Diskriminierung durch Kaufempfehlungssystem. Vor allem Betreiber brauchen hier ein „geschultes Auge", um Vorfälle in Einklang mit dem AI Act zu erkennen, zu bewerten und gegebenenfalls zu melden.

4.4 Fairness und Anbieter von KI-Systemen

Anbieter von KI-Systemen sind natürliche und juristische Personen, die KI-Systeme entwickeln oder entwickeln lassen, den Verwendungszweck bestimmen und/oder ein KI-System unter dem eigenen Namen verfügbar machen, siehe Artikel 3 (1) und Artikel 25 für Details. Ein Großteil der Pflichten im AI Act richtet sich an Anbieter, weil deren Entscheidungen während der Entwicklung ganz maßgeblich die Leistung und Sicherheit eines KI-Systems in der Nutzung beeinflussen. Im Beispiel der Kreditwürdigkeitsprüfung ist das Fintech-Startup der Anbieter.

Die Pflichten für Anbieter sind abhängig von der Risikoeinstufung des jeweiligen KI-Systems. Hochrisiko-KI-Systeme sind mit umfangreichen Anforderungen belegt, weil diese ein Konformitätsbewertungsverfahren durchlaufen müssen, bevor sie in Verkehr gebracht oder in Betrieb genommen werden dürfen (Artikel 43). Im Mittelpunkt stehen dabei die „Anforderungen an Hochrisiko-KI-Systeme" (Kapitel III, Abschnitt II, Artikel 8-15), wozu auch Risikomanagement, Data Governance, menschliche Aufsicht Genauigkeit, Robustheit und Cybersicherheit zählen. KI-Systeme mit begrenztem Risiko unterliegen sogenannten Transparenzpflichten (Artikel 50), die sicherstellen sollen, dass KI-basierte Inhalte oder interaktive Systeme (z. B. Chatbots) als solche erkennbar sind. Niedrigrisiko-KI-Systeme unterliegen keinen besonderen Pflichten, aber Anbieter werden ermutigt, einen freiwilligen Verhaltenskodex zu befolgen (Artikel 95).

In diesen Pflichten liegt ein zentraler Hebel, um mehr Fairness für betroffene Personen sicherzustellen. Die sorgfältige Auswahl von Trainingsdaten, Mechanismen für Transparenz und Reproduzierbarkeit sowie umfassende Informationen an die späteren Nutzer und Nutzerinnen des KI-Systems unterstützen den kontrollierten und verantwortungsvollen Einsatz von KI. Wie auch bei den anderen Stufen in der KI-Wertschöpfungskette gibt es auf der Ebene von Anbietern offene Punkte:

- Standards: Die Verwendung von Standards ist grundsätzlich freiwillig, aber laut Artikel 40 im AI Act bilden die sogenannten harmonisierten Standards die Grundlage für die Konformitätsvermutung, also die Konformität mit den „Anforderungen an Hochrisiko-KI-Systeme" (Artikel 8-15). Nach aktuellen Berichten (Stand August 2025) wird sich die Veröffentlichung der harmonisierten Standards deutlich verspäten, was auch deren praktische Umsetzung inklusive positiver Effekte für mehr Fairness in die Zukunft verschiebt.
- Marktüberwachung: Kompetente Behörden sind die Voraussetzung für die effektive Durchsetzung dieser Pflichten, jedoch ist mit Blick auf Umsetzungsgeschwindigkeit des AI Acts in den Mitgliedstaaten und die allgemein angespannte Haushaltslage unklar, wann und in welchem Umfang die jeweiligen Behörden bzw. deren Angestellte in angemessenen Umfang dafür befähigt und beauftragt sind.

4.5 Fairness und Anbieter von KI-Modellen mit allgemeinem Verwendungszweck

Anbieter von GPAI-Modellen sind natürliche und juristische Personen, ein KI-Modell mit allgemeinem Verwendungszweck (engl. „General Purpose AI Model") entwickeln oder entwickeln lassen oder ein GPAI-Modell unter dem eigenen Namen verfügbar machen, siehe Artikel 3 (1) und die Richtlinien des KI-Büros der EU Kommission für Anbieter von KI-Modell mit allgemeinem Verwendungszweck (vgl. AI Office 2025a). Die meisten bekannten Anbieter haben ihren Sitz außerhalb der EU, zum Beispiel Google DeepMind, Meta, Amazon, X oder Microsoft. Im Beispiel der Kreditwürdigkeitsprüfung könnte das Fintech Startup ein GPAI-Model als eine Komponente des spezifischen KI-Systems integrieren.

Die Pflichten im AI Act für Anbieter von GPAI-Modellen orientieren sich an der Leistungsfähigkeit des Modells und in Abhängigkeit, ob das Model „open source" oder proprietär ist. Die geringsten Pflichten gelten nach Artikel 53 (2) für „freie und offene" GPAI-Modelle, die eine Zusammenfassung ihrer Trainingsdaten veröffentlichen und europäisches Recht für Urheberrechte beachten müssen. Die nächste Stufe sind proprietäre GPAI-Modelle, die zusätzlich technische Dokumentation über Test- und Trainingsprozess veröffentlichen und relevante Informationen für nachgelagerte Anbieter (wie das Beispielhafte FinTech Startup) zur Verfügung stellen müssen. Die meisten Pflichten gelten für GPAI-Modelle mit systemischem Risiko, also Modelle, die über überdurchschnittliche Kapazitäten oder eine besonders hohe Reichweite verfügen. An-

bieter solcher Modelle müssen laut Artikel 55 zusätzlich einen Risikomanagementprozess umsetzen, Maßnahmen für Cybersecurity befolgen und einen Prozess für die Bearbeitung von Vorfällen betreiben.

All diese Pflichten zielen darauf ab, die Qualität und Robustheit von GPAI-Modellen zu verbessern, was einen positiven Effekt auf den letztendlichen Einsatz mit betroffenen Personen haben sollte, weil die Nachweise Sicherheit und Transparenz für nachgelagerte Anbieter schaffen. Am 02. August hat das KI-Büro den freiwilligen Verhaltenskodex für GPAI-Modelle (vgl. ebd. 2025b) veröffentlicht, in dem konkrete Umsetzungshinweise für Artikel 53 und 55 zu finden sind. Bevor diese Anforderungen einen Effekt zeigen, gilt es:

- Einhaltung: Zwar haben sich (Stand August 2025) mehrere große GPAI-Anbieter entschlossen, dem Kodex per Selbstverpflichtung zu folgen, dennoch steht aus, wie der Kodex konkret in der Praxis umgesetzt wird. Eine wesentliche Herausforderung dabei sind die unterschiedlichen Informationsbedarfe der zigtausend KI-Anwendungen bei nachgelagerten Anbietern, die auf die Auskünfte der GPAI-Anbieter angewiesen sind.
- Durchsetzung: Die Durchsetzung der Regeln für GPAI-Modelle wird teilweise vom KI-Büro übernommen (Artikel 75) und birgt zwei Herausforderungen. Erstens ist der Kodex freiwillig, sodass sich das KI-Büro im Zweifel auf Artikel 53 und 55 berufen kann, aber nicht auf den Kodex an sich. Zweitens hat die Durchsetzung eine wirtschaftliche und geopolitische Dimension, aufgrund der großen Marktanteile der US-amerikanischen GPAI-Anbieter.

5. Zwischenfazit

An dieser Stelle eignet sich ein Zwischenfazit, weil der Umsetzungsprozess des AI Acts in vollem Gang ist und es regelmäßig neue Fortschritte gibt. Mit Blick auf die eingangs gestellten Fragen zeigt dieser Beitrag, dass der AI Act konkrete Anforderungen an Fairness von KI-Systemen enthält und dass diese Anforderungen auf ethischen Prinzipien beruhen. Diese Beobachtung unterstützt die Ansicht, den AI Act als formalisierte Ethik zu bezeichnen. Die Anforderungen an Fairness verteilen sich über die verschiedenen Stufen der KI-Wertschöpfungskette: Betroffene Personen erhalten durch den AI Act neue Rechte (Recht auf Erläuterung, Recht auf Beschwerde), hingegen sehen sich Betreiber und Anbieter von KI-Systemen und GPAI-Modellen mit neuen Pflichten konfrontiert. Der Umfang der Pflichten folgt dabei stets dem Risikobasierten Ansatz: bei KI-Systemen ist die

Risikoeinstufung maßgeblich (Hochrisiko, begrenztes Risiko, niedriges Risiko) und bei GPAI-Modellen hängt es von der Klassifizierung mit oder ohne systemische Risiken ab.

Zwar stellen die neuen Rechte und Pflichten eine Tendenz zu „mehr Fairness in KI" in Aussicht, aber die tatsächliche Wirkung des AI Acts hängt von der Um- und Durchsetzung der neuen Regeln ab, wobei sich einige Herausforderungen stellen. Betroffene Personen können nur effektiv vor Diskriminierung geschützt werden bzw. sich davor schützen, wenn klar ist, wer genau damit gemeint ist und wenn diese Personen Kenntnis über ihre neuen Rechte im AI Act haben. Betreiber von KI-Systemen müssen interne Methoden und Fähigkeiten aufbauen. Einerseits, um potenzielle Diskriminierung im Rahmen von einer Grundrechtefolgenabschätzungen zu identifizieren und mitigieren, andererseits um Vorfälle während der Nutzung als solche zu erkennen und zu behandeln. Anbieter von KI-Systemen haben den „Löwenanteil" der Pflichten auf ihrer Seite und befinden sich gleichzeitig in einem Spannungsfeld. Die Pläne, sich durch KI einen Wettbewerbsvorteil zu verschaffen, werden durch Verzögerungen bei der Veröffentlichung der harmonisierten Standards und der Etablierung der nationalen Marktüberwachung gedämpft. Der Verhaltenskodex für GPAI-Modelle (Seit Juli 2025) ist eine wichtige Referenz für mehr Transparenz und klare Verantwortlichkeiten bei Basismodellen, aber seine freiwillige Natur birgt Herausforderungen im Bereich der Einhaltung von Anbietern und Durchsetzung von Behörden.

6. Ausblick

Es bleibt weiter spannend in der Arena um KI-Fairness, denn das Feld ist sehr dynamisch und viele Akteurinnen und Akteure, mit teilweise unterschiedlichen Vorstellungen, sind involviert. Die Umsetzung des AI Acts schreitet gemäß der Fristen in Artikel 113 voran und der 2. August 2025 war der jüngste Stichtag. Seither gelten zum Beispiel die Regelungen für die Nationale Aufsicht (Artikel 28), die Pflichten für Anbieter von GPAI-Modellen (Artikel 53 und 55), die Governance auf EU-Ebene (Artikel 64, 65) und die Sanktionen (Artikel 99). Der nächste große Meilenstein ist der 02. August 2026, wenn das Gros der Pflichten in die Anwendung kommt, insbesondere die Pflichten für bestimmte Hochrisiko-KI-Systeme sowie Transparenzanforderungen für KI-Systeme mit begrenztem Risiko.

Auf EU-Ebene gibt es verschiedene Handlungsstränge, die einen Einfluss auf die Fairness für KI-Systeme bzw. digitale Dienste im Allgemeinen haben können. Bereits 2024 hat die EU

Kommission einen Entwurf für den „Digital Fairness Act“ (vgl. EU Kommission 2024) präsentiert und sammelt bis zum 24. Oktober 2025 Feedback und Nachweise („evidence“) für die mögliche Wirkung, die ein solcher Act haben könnte. Das Ziel ist es, Verbraucherinnen und Verbraucher in der Digitalwirtschaft besser zu schützen. Parallel dazu arbeitet die EU Kommission an einem sogenannten „Omnibus Package“ (vgl. EU Kommission 2025) mit dem Ziel, bestehende Regulierung im digitalen Sektor zu vereinfachen („Simplification“), wobei es insbesondere um die Verschlankung und Integration geht, um langfristig die Wettbewerbsfähigkeit von Unternehmen zu stärken.

Fairness im Bereich KI kann zu einem „win-win“ werden, wenn die gängige Rhetorik a la „Innovation versus Regulierung“ abgelöst und durch „mehr Fairness führt zu mehr Innovation“ ersetzt wird. Umfrageergebnisse zeigen, dass sowohl Unternehmen als auch Privatpersonen KI unter anderem auch deshalb nicht nutzen, weil es Bedenken bezüglich der Zuverlässigkeit und Vertrauenswürdigkeit gibt. Sollte es auch mit Hilfe des AI Acts gelingen, die Qualität von KI-Systemen zu erhöhen, kann das die Adaption von KI vorantreiben. Als Resultat hieße das mehr Schutz für betroffene Personen und mehr Wettbewerbsfähigkeit und Produktivität in Unternehmen.

Bis dahin ist viel zu tun. Gemäß dem Sprichwort “it takes a village to raise a child” zeigt dieser Beitrag “it takes an AI value chain to get a fair AI system”. Es braucht Anstrengungen, Ideen und Zusammenarbeit, um eine faire digitale Zukunft in der EU zu erreichen. Der AI Act geht den Schritt von der Ethik zum Gesetz, aber es ist eher ein Marathon als ein Sprint, bis das Gesetz zur Realität wird.

Literaturverzeichnis

AI Office (2025a): Guidelines on the Scope of Obligations for Providers of General-Purpose AI Models Under the AI Act, URL: https://digital-strategy.ec.europa.eu/en/library/guidelines-scope-obligations-providers-general-purpose-ai-models-under-ai-act (aufgerufen am: 05/08/2025).

– (2025b): The General-Purpose AI Code of Practice, URL: https://digital-strategy.ec.europa.eu/en/policies/contents-code-gpai (aufgerufen am: 05/08/2025).

Atari, M. / Xue, M. J. / Park, P. S. / Blasi, D. E. / Henrich, J. (2023): Which Humans?, Cambridge: Havard University Press, URL: https://scholar.harvard.edu/sites/scholar.harvard.edu/files/henrich/files/which_humans_09222023.pdf (aufgerufen am: 05/08/2025).

Berti, L / Giorgi, F. / Kasneci, G. (2025): Emergent Abilities in Large Language Models: A Survey, URL: https://arxiv.org/abs/2503.05788 (aufgerufen am: 05/08/2025).

Bitkom (2024): Künstliche Intelligenz in Deutschland, URL: https://www.bitkom.org/sites/main/files/2024-10/241016-bitkom-charts-ki.pdf (aufgerufen am: 05/08/2025).

Danish Institute for Human Rights (2025): Human Rights Impact Assessment of Digital Activities, URL: https://www.humanrights.dk/publications/human-rights-impact-assessment-digital-activities (aufgerufen am: 06/08/2025).

EFI (2024): Gutachten zu Forschung, Innovation und technologischer Leistungsfähigkeit Deutschlands 2024, Berlin: EFI.

EU Kommission (2019): Ethics Guidelines for Trustworthy AI, URL: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai (aufgerufen am: 05/08/2025).

– (2024): Digital Fairness Act, URL: https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/14622-Digital-Fairness-Act_en (aufgerufen am: 05/08/2025).

– (2025): Commission Proposes to Cut Red Tape and Simplify Business Environment, URL: https://commission.europa.eu/news-and-media/news/commission-proposes-cut-red-tape-and-simplify-business-environment-2025-02-26_en (aufgerufen am: 07/08/2025).

Europäische Union (2000): Charta der Grundrechte der Europäischen Union, URL: https://www.europarl.europa.eu/charter/pdf/text_de.pdf (aufgerufen am: 05/08/2025).

Federal Anti-Discrimination Agency (2019): Risks of Discrimination through the Use of Algorithms, URL: https://www.antidiskriminierungsstelle.de/EN/homepage/_documents/download_diskr_risiken_verwendung_von_algorithmen.pdf?__blob=publicationFile&v=1 (aufgerufen am: 05/08/2025).

Future of Life Institute (2025): AI Safety Report 2025, URL: https://futureoflife.org/wp-content/uploads/2025/07/FLI-AI-Safety-Index-Report-Summer-2025.pdf (aufgerufen am: 06/08/2025).

ISO / IEC (2021): ISO/IEC TR 24027:2021 Information Technology – Artificial intelligence (AI) – Bias in AI Systems and AI Aided Decision Making.

Kaiser, C. / Buder, F. / Biró, T. (2024): ChatGPT und Co. im Alltag: Nutzung, Bewertung und Zukunftsvisionen. Ein Drei-Länder-Vergleich, NIMpulse 7, Nürnberg: Nürnberg Institut for Market Design.

KitaMatch (2023): Startseite, URL: https://kitamatch.com/ (aufgerufen am: 06/08/2025).

Norori, N. / Hu, Q. / Aellen, F. M. / Faraci, F. D. / Tzovara, A. (2021): Addressing Bias in Big Data an AI for Health Care: A Call for Open Science, in: Perspective, Jg. 2 / Nr. 10, Artikel 100347.

Rughinis, R. / Rughinis, C. / Vulpe, S. N. / Rosner, D. (2019): From Social Netizens to Data Citizens: Variations of GDPR awareness in 28 European countries, in: Computer Law & Security Review, Jg. 42, Artikel 105585, DOI: 10.1016/j.clsr.2021.105585.

Smuha, N. A. (2024): The Work of the High-Level Expert Group on AI as the Precursor of the AI Act, in: Ceyhun, N. P. / Forgó, N. / Valcke, P. (Hrsg.): AI Governance and Liability in Europe – A Primer, URL: https://ssrn.com/abstract=5012626 (aufgerufen am: 07/08/2025).

Startup Verband (2024): Startups und Generative KI – Ein neues Zeitalter beginnt, URL: https://startupverband.de/fileadmin/startupverband/forschung/studien/ki/Startups_Generative_KI_2024.pdf (aufgerufen am: 07/08/2025).

The Concept of Corporate Digital Responsibility (CDR)

In the Light of a Philosophical Analysis of the Concept of Responsibility in the Age of AI

Christian Hugo Hoffmann

1. Introduction

In the following, I bring forward some key insights from my recent book "Zukunftsethik der Künstlichen Intelligenz: Perspektiven und Strategien für ein verantwortungsvolles und wirtschaftliches Handeln von morgen" (cf. Hoffmann 2025) and apply them to the discourse on Corporate Digital Responsibility (CDR). For the latter, I particularly refer to a paper by two researchers in business administration at a German university of applied science, Knopf and Pick (2023), since the academic discourse seems to be coined by persons that are closer to, on the one hand, business and social sciences than to the humanities and philosophy and to, on the other hand, the business practice than the ivory tower.

Using this example, the present paper pursues the goal of outlining some flaws in the discourse on CDR which result from a lack of rigorous philosophical analysis. This finding will then allow us to derive lessons on how to improve the future discourse on CDR.

2. *Setting the Scene: Approach by Knopf and Pick*

The authors kick-off their investigation by observing that

> [t]he digital transformation of the economy and society is fundamentally changing the possible value creation processes of companies. These changes increasingly involve the processing of personal data into information and knowledge that can be used to create individualised user experiences or gain insights from aggregated data. Customer relationships and the interconnectedness of customers are similarly changing as a result of this digital transformation. As customers become more demanding and involved, voluntary activities that go beyond compliance and regulation are becoming increasingly important. At the same time, companies are increasingly recognising the potential of privacy-friendly goods and services. They are using the heterogeneity of their offerings to develop unique selling propositions that focus on fulfilling their digital responsibilities (e.g., Apple, Samsung). In this way, digital responsibility fosters innovation in goods, services and digital markets. The voluntary responsibility of companies in dealing with digital technologies and creating value from data is discussed under the term "corporate digital responsibility" (CDR) (Knopf/Pick 2023: 469).

Knopf and Pick acknowledge that this young field of research roots in the study of Corporate Social Responsibility (CSR), which has been around for decades, yet they propose to regard CDR as a discipline in its own right. The authors attempt to back up the latter claim by conducting a systematic literature synthesis. Their paper thereby puts emphasis on

> definitional approaches, approaches to describing the entrepreneurial motivation for CDR activities and their consequences for corporations, their business relationships and society (ibid.).

The core element of their paper are the following two tables where the first presents 18 definitions of the term "CDR" and the second evaluates them along ten criteria that the authors simply stipulate.

Authors	Definitions
BMUV (2021)	"CDR is a voluntary corporate activity, particulary considering the consumers' persepctive, which strives to go beyonf what is required by law to shape the digital world for the advancement of society"
Carl et al. (2022)	"CDR activities exceed the legally binding (national) minimum requirements and rather describes the voluntary accaptence of additional responsibilities"
Dörr et al. (2021)	"A set of practices and behaviours that help an organisation to use data and digital technology in ways that are preceived as socially, economically and envirnomentally responsible"
Driesens et al. (2017)	"CDR is a voluntary commitment. It starts with the need to conform to legal requirements and standards – for handling customer data, confidential, intellectual property and so on – but it also extends to wider ethical considerations and the fundamental values that an organization operates by"
Elliot et al. (2021)	"A potential collaborative mechanism to navigte such complexity, proposing guidance frameworks towards responsible corporate digital action in preserving societal interets"; "A voluntary commitment by organizations fullfilling the corporate rationalizers' role in representing community interests to inform "good" digital corporate actions and digital sustainability (i. e. data and algorithma) via collaborative guidance on adressing social, economic and ecological impacts on digital society"
Herden et al. (2021)	"Is an extension of a firm's responsibilites which takes into account the ethical opportunites and challenges of digitalization"
Huber (2022) (translated from german)	"Framework for action and decision-making that shows how AI risks and opportunities are to be balanced in order to meet social responsibility of companies with regard to digital issues"; "Perception of a voluntary, beyond the legally binding, corporate responsibility for the consequences of the development, disseminatio or use of digital technologies on society as a whole, but also on the concrete individual person"
Kenning (2020) (translated from german)	"Extent […] to which companies and organizations use digital technologies to achieve societal goals in the areas ecology, economy and social issues, as well as the extend to which companies and organizations take into account aspects of data protection and data security as well as use data and digital technologies responisble"
Leonard (2016)	"Expanding the remit of CSR to address the impact of the digital tools and environments that business operate in"
Lobschat et al. (2021)	"The set of shared values and norms guiding an organization's operations with respect to the creation and operation of digital technology and data"
Mihale-Wilson et al. (2021, 2022)	"CDR seeks to ensure an ethical and responsible development, deplyoment and use of digital technologies and data"; "Part of the "obligations"that companies have toward society"
Schymura (2018) (translated from german)	"Corporate Digital Responsibility is an understanding of corporate responsibility in and for a digital society. It involves a regulated and a voluntary level: on the one hand, the observance of relevant laws or derectives, on the other hand, the exercise of a voluntary responsibility in shaping the digital society"
Suchacka (2019)	"CDR means the awareness of duties binding the organisations active in field of technological development and using technologies to provide services"
Thelisson et al. (20219)	"[…] means a kind of digital corporate social responsibility"
Trittin-Ulbrich/Böckel (2022)	"CDR emphasizes the voluntary, self-regulatory character of coporate commitment to responsible digital innovation"
Van der Merwe/Al Achkar (2022)	"As the set of practices, policies and governance structures of coparations as they relate to the digital transformation"
Weißemberger/Marrocco (2022)	"CDR is a voluntary corporate orientation to ensure a responsible use of digital technologies"
Wirtz et al. (2023)	"We define CDR in the context of serivice as the principle underpinning a service firm's ethical, fair and protective use of data technology when engaging with customers within their digital service ecosystem"

FIGURE 1: DEFINITIONS OF CORPORATE DIGITAL RESPONSIBILITY (SOURCE: KNOPF/PICK 2023: 472)

Categories	Scope of Responsibilitx				Scope of Activites			Scope of Beneficaires		
	Eco-nomic	Legal	Ethi-cal	Philan-thropic	Envi-ron-mental	Social	Govern -ance	Cor-porate	Corpo-rate Relations	So-ciety
Sum	2	11	14	5	4	8	10	3	3	6
BMUV (2021)		X	X	X		X		X	X	X
Carl et al. (2022)		X	X							
Dörr et al. (2021)	X		X		X	X	X	X		
Driesens et al (2017)		X	X	X		X				X
Elliot et al. (2021)	X	X	X	X	X	X	X			X
Herden et al. (2021)			X							
Huber (2022)		X	X			X	X	X	X	X
Kenning (2020)		X			X	X	X			
Leonard (2016)					X		X			
Lobschat et al. (2021)			X				X			
Mihale-Wilson et al. (2021, 2022)		X	X			X	X			X
Schymura (2018)		X	X	X						X
Thelisson et al. (20219)						X				
Trittin-Ulbrich/Böckel (2022)			X	X						
Van der Merwe/Al Achkar (2022)		X					X			
Weißemberger/ Marrocco (2022)			X							
Wirtz et al. (2023)		X	X				X		X	

FIGURE 2: CDR DEFINITIONS BY SCOPE OF RESPONSIBILITY, ACTIVITIES AND BENEFICIARIES (SOURCE: KNOPF/PICK 2023: 473)

Based on that, Knopf and Pick (2023) identify the research gap of a lack of consensus on a unified CDR definition, and assert that a better understanding of CDR would be achieved by a unification of the different proposals (cf. ibid.: 474), which however looks premature in the light of the subsequent key findings from a clear philosophical exploration of the concept of responsibility in the context of AI, entrepreneurship and effective ethics (cf. Hoffmann 2025).

3. Key Insights for Clarification

According to Hoffmann (2025), the concept of moral responsibility, which in turn is crucial in the debate on technology ethics, can be systematized as follows:

	(1)	**(2)**	**(3)**
(A) WHO is responsible (Agent)	Individual	Corporation	Society
(B) WHAT	Action	Omission	Product
(C) WHOM to (patient)	Individual	Collective	Thing
(D) WHEREFORE	Moral rules	Social values	State laws
(E) WHAT OF	Conscience	Judgment of others	Court
(F) WHEN	After: prospective	Momentarily	Before: retrospective
(G) HOW	Active	Virtual	Passive
(H) HOW good or bad	Positive	Neutral	Negative
(I) WHAT scope of the consequences of the action (in terms of location / time)	Unremarkable	Cumulative	Irreversible

FIGURE 3: MORPHOLOGICAL MATRIX OF RESPONSIBILITY TYPES (PARTIAL) (SOURCE: ROPOHL 1994: 112).

Referring to these questions, a suitable systematization of the concept of responsibility can be proposed as a morphological matrix of responsibility types in Figure 3, whereby (A) is aimed at the acting subject, (C) represents the opposite side, i.e. the person to whom the action happens, (D) represents a central element of the concept of responsibility, since otherwise it would be unclear why one operates with a concept of responsibility at all and (E), closely linked to (D), stands for the instance vis-à-vis which one is held responsible; (B) is implied because responsibility

is based on acting subjects; (F) corresponds with the time dimension and (H) with the evaluation dimension of the consequences of actions and (G) was introduced by Ropohl (1994) as a new additional differentiation option. Dimension (I), or the last-mentioned question in Figure 3, was first recorded in Hoffmann (2025), as AI is essentially interwoven with its object of investigation. We inhabit an AI world and witness an era of AI, and AI makes it necessary to include new ethical aspects.

With the help of this matrix, individual types of responsibility relevant to Hoffmann's (2025) study can be outlined and specified. Concepts of responsibility, such as those underlying the idea of AI valuation and responsible tech entrepreneurship, can be reconstructed along this scheme (Figure 3) and the following lines:

(A) Agent(s):	(A1): AI developer (researcher, product manager, …)
	(A1): AI user
	(A1): AI system
	(A2): AI company / companies
	(A3): Modern societies in which AI occurs
(B) Responsible for (B1-3):	Ostensibly actions (B1) that lead to AI systems or AI products, for example, but also the AI products themselves Or omitted actions that would be desirable
(C) Patient(s):	(C2): AI (strikingly often) affects many, not individuals
(D) Evaluation system (D1-2):	Some concepts of responsibility are shaped by value-ethical approaches, others by moral rules and norms; however, since the ethicists typically deal with moral responsibility, laws are ruled out
(E) Instance	(E2): Primarily the judgment of others
(F) Time dimension	(F1): In this respect, building on the findings from Hoffmann (2025), a plea is made for forward-looking responsibility

(G) Involved	(G1): Analogous to (F), although AI specifics and the dynamic complexity of our social systems allow for the whole variety, i.e. also (G2) and (G3)
(H) Consequences of action	(H1 and H3): Technology assessment (especially in the spirit of Jonas, 1979/1984) is classically aimed at threats and thus H3, whereas Hoffmann (2025) with its focus on entrepreneurship and start-ups is oriented towards innovations and opportunities and thus H1
(I) Scope	(I2–3): AI technology changes the consequences of actions in such a way that irreversibility and cumulativeness occur

Hoffmann (2025) then zooms into each row or category and scrutinizes to what extent responsibility can and should be born and what restrictions are to be singled out. Thereby, he pays special attention to the limits of responsibility concepts. In terms of the latter, examples abound to motivate how AI poses problems to operate with the concept of responsibility. The phenomenon that perhaps stands out most is the so-called "responsibility gap" which results from the use of AI, and which can be illustrated as follows:

For example, if I (a human person) compete against the South Korean professional Lee Sedol in the board game "Go" and lose, then I am responsible for this failure (responsibility with a negative connotation, H3) because I thought about my moves and executed them to the best of my knowledge and free from constraints. However, who or what should be praised for the success (responsibility with a positive connotation, H1) of the AI AlphaGo in 2016, which "beat" Lee Sedol as a computer program? The developers and programmers at DeepMind? The entire company, right down to the janitor? The person who moved the board pieces on behalf of AlphaGo? AlphaGo itself? We must first clarify these questions before we as a society can establish clear and transparent rules for the responsible use of AI. Examples such as this one pose the question of responsibility forcefully and puzzle over the answers. Indeed, Matthias (2004 and many others since, e.g. Kiener 2022) argues that there is a growing responsibility gap: The more complex computing and AI technologies become, and the less humans can directly control or intervene in the behavior of these technologies, the less we can reasonably hold humans responsible for these technologies or their use, respectively. Despite the ongoing philosophical debates on this topic, most analyses of moral responsibility exhibit at least the following three conditions:

1. There should be a causal link between the agent and the outcome of her actions. The moral subject is usually only held responsible if they had at least some control over the outcome of the events. How close the causal chains must be is a point of contention in the expert debate and a position is taken in Hoffmann (2025) following Santoni and van den Hoven (2018).
2. The agent must know and be able to weigh up the possible consequences of their actions. We tend to absolve someone of blame (e.g. young children) if she or he could not have known that her or his actions would lead to a harmful event.
3. The agent must be able to freely decide to act in a certain way. Responsibility is the reverse side of freedom, as it were. In other words, it makes no sense to hold someone responsible for suffering or, conversely, positive benefits, i.e. (un)desired consequences of action, if her or his actions were completely determined by external forces. Or to put it in a crisp, thanks to Kant familiar formula that can be applied here: Should implies can.

A closer look at these three requirements or conditions for responsibility concepts shows that, on the one hand, the dynamic complexity of our current socio-economic systems (cf. Forrester 1961; Ulrich/Probst 1990) and, on the other hand, computer technology or, more specifically, AI, which are arguably a concrete manifestation of dynamic complexity, can call into question the fulfillment of each of these conditions. Behind the scenes, what we observe in our dynamically complex and AI-interwoven world is as follows:

Ad 1: As seen in the case study above, AI can obscure the causal relationships between a person's actions and the possible consequences of those actions. In particular, the involvement of a large number of actors in the development and use of AI technologies, sometimes with a large physical and/or temporal distance between the respective actions and the associated consequences of the actions, leads to the so-called problem of "many hands" (cf. Nissenbaum 1994; Doorn/van de Poel 2012): As traced in the case study above with AlphaGo, it is difficult to determine who is responsible for what when several people have contributed to the outcome of the events. Several AI engineers and their managers were involved in the success of AlphaGo; however, from a purely causal perspective, the person who mechanically moved the game pieces in the name of AlphaGo was also involved.

Ad 2: According to the second condition, it would be inappropriate to hold someone responsible for something if he or she could not have known at the time of the action that his or her actions could lead to harm or suffering. Although AI technologies can increase the predictability

and calculability of the consequences of actions – think of AI-supported climate models, for example – the reverse is also true: AI often makes it more difficult to understand the outcomes of decisions; "often" in the sense that powerful deep learning tools dominate the AI world these days, which in turn represent black boxes to a particular extent – although the feature of the black box or this problem of opacity is much greater and more fundamental. The following example of the difficulty of assessing the consequences of actions by AI confirms this finding: AlphaGo and its successor AlphaGo Zero have reached a world championship level in the game of Go, in that the AI programs have chosen moves that neither their developers nor experts such as Lee Sedol could predict (cf. Silver/Hassabis 2016; Silver et al. 2017; Tegmark, 2018: 88). When analyzing AlphaGo's game, the focus was not only on its victory over the South Korean professional, but also on supposed "AI creativity".

Ad 3: Freedom of action is probably not only the most important condition for the attribution of moral responsibility, but it is also arguably one of the most controversial (cf. Talbert 2019). The community of moral subjects tends to absolve people of moral blame when they had no choice but to act as they did. This is because the importance of freedom of action, which is usually associated with a gradual concept (e.g. an adult is, ceteris paribus, freer to act than a child) for the attribution of moral responsibility is based on the fact that free, self-determined action expresses the authentic motivation, motives and intentions of the moral subject.

Like other technologies, AI adds an additional layer of complexity when it comes to determining whether someone is capable of acting freely, as AI affects the decisions that decision-makers make and how they make them. One of the most important AI application areas is the automation and control of decision-making processes. A seminal example of how actors become less free through the use of AI is Amazon's now defunct AI recruitment tool (cf. Chang 2023).[1]

---

[1] Amazon has shut down its experimental recruitment tool based on AI after it was found to discriminate against women. The company had developed the AI to scour the internet and recognize potentially suitable applicants and rate them with one to five stars. However, the algorithm learned to systematically devalue the CVs of women for technical professions such as software development. Machine learning algorithms like the one used by Amazon are trained to recognize patterns in large data sets in order to predict outcomes. In Amazon's case, the algorithm used all the CVs submitted to the company over a ten-year period to learn how to recognize the best applicants. Given the low proportion of women in the company, as is the case in most technology companies, the algorithm quickly recognized the male dominance and considered it a success factor, which in philosophy is recognized as a fallacy where

To summarize: All three (necessary) conditions for assuming responsibility are called into question by AI. What results from this realization? Is it still possible to operate meaningfully with "responsibility" in the age of AI? Are there gaps in responsibility here and there? What solution strategies are available?

There are solution strategies for satisfactorily answering the following two questions: Who exactly is to be held responsible if the AI machine makes/will make a mistake? Or who takes the credit when the machine successfully solves/will solve a problem? Five such strategies are briefly named here and are explained and discussed by Hoffmann (2025), before the fifth strategy is highlighted as the most promising.

- Strategy 1: AI bears responsibility (not compelling according to Hoffmann 2025)
- Strategy 2: Responsibility gaps do not need to be closed at all (does not correspond to our common practice)
- Strategy 3: Moral actors voluntarily assume responsibility (possible and desirable, but lacks incentivization)
- Strategy 4: Semantic differentiation in the attribution of responsibility (interesting, yet limited in terms of impact according to Hoffmann 2025)
- Strategy 5: System view by focusing on human-machine teams (most promising according to Hoffmann 2025).

In light of these and more findings, Hoffmann (2025) proposes the subsequent responsibility matrix when applied to nowadays AI:

causality is confused with correlation. Because the AI was now using the results of its own predictions to improve its accuracy, despite Amazon's leadership in AI technology, it got stuck in the trap of sexism towards female applicants, meaning the AI inherited undesirable human traits such as bias and discrimination that have been a problem in recruitment for years. The human decision-makers in Amazon's HR department thus became less free in their recruiting decisions while using their AI tool, as many actually suitable female candidates were wrongly and unthinkingly not looked at and the HR professionals thus unknowingly missed their goal to attract the best talent for Amazon.

(A) WHO is responsible (Agent)	Lexical order, which is specified in Hoffmann (2025): ▪ AI companies (as sufficiently large, stable systems appropriate to absorb the complexity of AI and companies are concrete actors) ▪ AI developers (concrete actors who have a lot of influence on AI) ▪ AI users (specific, but disadvantaged actors in AI markets, structural information asymmetry) ▪ Civil society (no specific actor, whereas responsibility requires a clear address for attribution) ▪ Moderate regulator ("the last resort")
(B) WHAT	Focus on actions
(C) WHOM to (patient)	Collective
(D) WHEREFORE	"Plurality of orientation systems"
(E) WHAT OF	The judgment of others
(F) WHEN	Primary: prospective Secondary: retrospective
(G) HOW	Primary: active Secondary: virtual and passive
(H) HOW good or bad	Primary: positive Secondary: negative
(I) WHAT scope of the consequences of the action (in terms of location / time)	Cumulative and irreversible

FIGURE 4: MORPHOLOGICAL MATRIX OF RESPONSIBILITY TYPES REVISITED (SOURCE: HOFFMANN 2025 [FORTHCOMING])

4. Limitations of the Analysis and Implications for the Future Discourse on CDR

When we now bring together the results from sections 2 and 3 of the present chapter, it has to be noted first that there is a certain discrepancy. While Knopf and Pick (2023) as well as the term CDR refer to digitization, Hoffmann (2025 [forthcoming]) is investigating AI which is not the same. My understanding is that AI is a form of automation which presupposes digitization. (To arrive at a deeper understanding of AI which also considers what the I in AI stands for, the reader is referred to Hoffmann 2022) Moreover, and in particular, I would hold that the findings by

Hoffmann (2025 [forthcoming]) do not only apply to AI, but also to digitization. With this in mind, we can assess and comment on the paper by Knopf and Pick (2023), and potentially the wider debate on CDR as follows:

- Knopf and Pick (2023) do not problematize the concept of responsibility at all, at least not in an analytic philosophical fashion.
- If one was willing to defend their analysis on the grounds that the authors differentiate between an economic, legal, ethical and philanthropic dimension of responsibility (see Figure 2), then a philosopher must respond that this hardly does justice to a proper philosophical or conceptual analysis since the problem of what is meant is simply moved by one level.
- The veritable pitfalls that operating with a responsibility notion pose in connection with AI or digitization for that matter are disregarded.
- Their readers are not made aware of how the concept of responsibility works, i.e., the crucial elements of the concept of responsibility as outlined, for example, in Table 3, are not taken into account. This provokes misunderstandings.
- Knopf and Pick (2023) scratch the surface at most when they distinguish between a legal use or context of responsibility and a use of the term of responsibility that goes beyond that. To philosophers such as Hoffmann (2025 [forthcoming]), this point is more regarded as a starting point of a study than to be taken as a genuine result.
- None of the 18 definitions the authors reviewed and assembled in Table 1 are inspired by philosophy, its rigor and objective of conceptual clarity. For this reason, it is doubtful to what extent these definitions are helpful at all; well-exemplified by the proposal no. 14 in the table: "[CDR] means a kind of digital [merely spelling out the "d" in "CDR"] CSR". In IT, this would well be labelled as a case of garbage-in-and-garbage-out.
- The issue as they claim is "not" a lack of consensus on a unified CDR definition or a better understanding of CDR through a unification of the different proposals from Table 1; in fact, this might do more harm than good to build master definitions and consensus on top of little conceptual understanding. The way forward ought to take the complexity of the notion of responsibility, including its challenges seriously and respond to them first.

The reader is invited to learn more about how AI ethics should not be conceived of as a regulatory and interventionist technology ethics, but rather as innovation-friendly business ethics that

strengthens entrepreneurship. We, therefore, point the interested reader to the full book "Zukunftsethik der Künstlichen Intelligenz: Perspektiven und Strategien für ein verantwortungsvolles und wirtschaftliches Handeln von morgen" (ibid.)

References

Chang, X. (2023): Gender Bias in Hiring: An Analysis of the Impact of Amazon's Recruiting Algorithm, in: Advances in Economics Management and Political Sciences, Vol. 23 / No. 1, 134–140.

Doorn, N. / van de Poel, I. (2012): Editors Overview: Moral Responsibility in Technology and Engineering, in: Science and Engineering Ethics, Vol. 18 / No. 1, 1–11.

Forrester, J. W. (1961): Industrial Dynamics, Cambridge: MIT Press.

Hoffmann, C. H. (2022): The Quest for a Universal Theory of Intelligence. The Mind, the Machine, and Singularity Hypotheses, Berlin: De Gruyter.

– (2025): Zukunftsethik der Künstlichen Intelligenz: Perspektiven und Strategien für ein verantwortungsvolles und wirtschaftliches Handeln von morgen, Baden-Baden: Karl Alber [forthcoming].

Jonas, H. (1979/1984): Das Prinzip Verantwortung: Versuch einer Ethik für die technologische Zivilisation, Frankfurt am Main: Suhrkamp.

Kiener, M. (2022): Can we Bridge AI's responsibility gap at Will?, in: Ethical Theory and Moral Practice, Vol. 25 / No. 4, 575–593.

Knopf, T. / Pick, D. (2023): Corporate Responsibility for Digital Innovation: A Systematic Review of the Literature. Proceedings of the 18th European Conference on Innovation and Entrepreneurship, ECIE 2023, URL: https://papers.academic-conferences.org/index.php/ecie/article/download/1601/1585/6348 (accessed: 14/02/2025).

Matthias, A. (2004): The responsibility Gap: Ascribing Responsibility for the Actions of Learning Automata, in: Ethics and Information Technology, Vol. 6 / No. 3, 175–183.

Nissenbaum, H. (1994): Computing and Accountability, in: Communications of the Association for Computing Machinery, Vol. 37 / No. 1, 72–80.

Ropohl, G. (1994): Das Risiko im Prinzip Verantwortung, in: Ethik und Sozialwissenschaften, Vol. 5 / No. 1, 109–120.

Santoni de Sio, F. / van den Hoven. J. (2018): Meaningful Human Control over Autonomous Systems: A Philosophical Account, in: Front Robot AI, Vol. 5 / Article 15, 5–15.

Silver, D. / Hassabis, D. (2016): AlphaGo: Mastering the ancient game of Go with Machine Learning. Google AI Blog, URL: https://ai.googleblog.com/2016/01/alphago-mastering-ancient-game-of-go.html (accessed: 31/01/2025).

Silver, D. / Schrittwieser, J. / Simonyan, K. / Antonoglou, I. (2017): Mastering the Game of Go Without Human Knowledge, in: Nature, Vol. 550 / No. 7676, 354–359.

Talbert, M. (2019): Moral Responsibility, in: Zalta, E. N. / Nodelman, U. (eds.): The Stanford Encyclopedia of Philosophy, URL: https://plato.stanford.edu/entries/moral-responsibility/ (accessed: 25/02/2025).

Tegmark, M. (2018): Life 3.0. Being Human in the Age of Artificial Intelligence, London: Penguin.

Ulrich, H. / Probst, G. J. B. (1990): Anleitung zum ganzheitlichen Denken und Handeln. Ein Brevier für Führungskräfte, Bern/Stuttgart: Haupt.

Gemeinsamkeiten bei Corporate Digital Responsibility schaffen[1]

Integration von Erkenntnissen, Definitionen und Expert*innenperspektiven

K. Valerie Carl, Marie Blachetta, Beatriz Bilfinger, Alexander Brink, Annette Cerulli-Harms, Anna-Sophia Christ, Karin Dietl, Saskia Dörr, Marc P. Hauer, Maria Hinz, Jens-Rainer Jänig, Katharina Keilhofer, Christian Kurtz, Cristina Mihale-Wilson, Benjamin Müller, Bartosz Przybylek, Nicole Schaufler, Isabelle Karoline Schlegel, Maike Scholz, Felix von Roesgen und Jakob Wößner

1. Warum eine gemeinsame Definition von Corporate Digital Responsibility Wirkung erzielt

Die digitale Transformation birgt eine Vielzahl von Chancen für verschiedene Stakeholder*innen (z. B. Verbraucher*innen, Unternehmen, Gesellschaft). Unternehmen können ihre Prozesse, Produkte und Dienstleistungen gezielt verbessern (vgl. Benbya et al. 2021; Kordzadeh/Ghasemaghaei 2022). Verbraucher*innen können u. a. von einer allgegenwärtigen Unterstützung in ihrem Alltag profitieren – etwa durch immer fortschrittlichere intelligente Wohnumgebungen, intelligente Autos und persönliche digitale Assistenten (z. B. Keller 2020; Mädche et al. 2019; Santos et al. 2016).

---

[1] Die Veröffentlichung ist zuerst als Denkimpuls bei der Initiative D21 erschienen und basiert auf dem Paper „Establishing Common Ground in Corporate Digital Responsibility: Integrating Understandings, Definitions, and Expert Perspectives", das unter der Creative Commons BY-SA 4.0 in der Informatik 2025 veröffentlicht wird: 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 [im Erscheinen]). Für den vorliegenden Abdruck wurde der Beitrag an die Zitations- und Formatvorgaben dieses Bandes angepasst; sie können daher leicht vom Original abweichen. Die Autor*innen danken Assistant Professor Dr. Lara Lobschat (Universität Maastricht, Niederlande) für ihre wertvollen Anregungen und ihr Feedback.

Die Gesellschaft als Ganzes kann durch die Entwicklung von unterstützenden Anwendungen, Diensten und Produkten, die durch die fortschreitende Digitalisierung ermöglicht werden, potenziell von mehr Nachhaltigkeit im ökologischen oder sozialen Sinne profitieren (z. B. Brîncoveanu et al. 2024).

Diese Chancen gehen jedoch für alle Stakeholder*innen auch mit Risiken einher. Diskriminierende, auf Künstlicher Intelligenz (KI) basierende Systeme, eine zunehmende Bedrohung der Privatsphäre und der Datensicherheit sowie psychische Risiken zählen zu den am meisten diskutierten Herausforderungen im Zusammenhang mit der Digitalisierung (z. B. Mikalef et al. 2022; Ransbotham et al. 2016; Tarafdar et al. 2013; Vial 2019; Wong et al. 2021). Die nachteiligen Auswirkungen treffen die Akteure auf unterschiedliche Weise: Wenn Unternehmen diskriminierende KI-basierte Systeme nutzen, können sie einen Vertrauensverlust bei ihren Kund*innen erleiden, was zu gravierenden wirtschaftlichen Folgen führen kann. Verbraucher*innen stehen hingegen häufig vor der schwierigen Abwägung, ob sie für den Mehrwert fortschrittlicher digitaler Dienste, Produkte und Technologien immer mehr persönliche Daten zur Verfügung stellen oder ihre Privatsphäre schützen sollten (vgl. Adjerid et al. 2018; Bélanger/Crossler 2011).

Verstöße gegen Datenschutz und Datensicherheit können Verbraucher*innen wirtschaftlich und psychisch belasten und sind daher ein zentrales Anliegen im digitalen Zeitalter. Um potenziell schwerwiegende Folgen wie Imageverlust, verändertes Verbraucher*innenverhalten oder rechtliche Konsequenzen zu vermeiden, gehen Unternehmen daher zunehmend freiwillige Verantwortungszusagen ein (vgl. Carl et al. 2023). Dadurch gewinnt das Konzept der Corporate Digital Responsibility (CDR) sowohl in der Praxis als auch in der wissenschaftlichen Debatte an Aufmerksamkeit.

Parallel dazu ist eine Vielzahl unterschiedlicher (praktischer) Rahmenwerke und (wissenschaftlicher) Publikationen entstanden. Während sich zunehmend ein Konsens über das allgemeine Verständnis von CDR abzeichnet (vgl. Mihale-Wilson et al. 2022), fehlt weiterhin eine übergeordnete Definition, die die verschiedenen Perspektiven der letzten Jahre bündelt. Eine solche Standarddefinition könnte die wissenschaftliche Debatte über CDR voranbringen, weil sie die Debatte vom Konzept selbst wegführt und den Blick stärker auf seine empirische Untersuchung und Operationalisierung lenkt. Damit käme die Forschung dem häufig formulierten Aufruf nach, sich von rein theoretischen Überlegungen zu lösen (vgl. ebd.; Müller 2022). Eine gemeinsame Definition nützt auch der Praxis: Ein solches gemeinsames Verständnis erleichtert den Einstieg

und senkt gerade für Nicht-Expert*innen die Einstiegshürde beim Zugang zum Thema CDR. Dies könnte mehr Unternehmen dazu motivieren, sich mit CDR zu beschäftigen und das Konzept in der Praxis zu operationalisieren, was letztlich zu einer größeren Verbreitung der verantwortungsbewussten Digitalisierung im Allgemeinen führen würde.

Mit diesem Denkimpuls wollen wir Kräfte aus Praxis, Forschung, Zivilgesellschaft und öffentlichen Institutionen in Deutschland, Österreich und den Niederlanden bündeln, um eine gemeinsame Definition von CDR zu entwickeln, auf die sich Expert*innen aus der CDR-Community einigen können. Wir haben dafür einen iterativen Ansatz gewählt, der auf verschiedenen bestehenden Definitionen und Rahmenwerken von CDR aufbaut.

2. Die Entwicklung von Corporate Digital Responsibility

CDR ist ein vergleichsweise junges Konzept. Während das verwandte Konzept Corporate Social Responsibility (CSR) – also der sozialen Verantwortung von Unternehmen – seit Jahrzehnten etabliert ist, wird CDR erst etwa 2015 erstmals erwähnt. In diesem Zeitraum wurde das Konzept zum Beispiel von der Unternehmensberatung Accenture in die Öffentlichkeit getragen (vgl. Cooper et al. 2015). Zunächst fand die Idee vor allem in der Praxis Beachtung und wurde von den ersten Unternehmen in deren Aktivitäten und Strategien integriert. Im Mai 2018 gründete das damalige Bundesministerium der Justiz und für Verbraucherschutz die CDR-Initiative, in der sich einige Pionierunternehmen der CDR zusammengeschlossen haben, um als Vorbilder und Lernplattform für andere Unternehmen zu dienen und das Thema sichtbar zu machen. Seit 2021 prämiert zudem der CDR-Award Unternehmensprojekte, die digitale Verantwortung besonders nachhaltig oder umfassend in der Unternehmenspraxis verankern.

In der Praxis gibt es heute bereits eine Vielzahl unterschiedlicher CDR-Aktivitäten und -Initiativen, was die zunehmende Praxisrelevanz des Themas unterstreicht – auch wenn es noch nicht flächendeckend umgesetzt ist. In der Forschung hingegen rückte CDR erst im Jahr 2021 stärker in den Fokus (z. B. Herden et al. 2021; Lobschat et al. 2021; Mihale-Wilson et al. 2022). Seitdem findet immer mehr wissenschaftliche Auseinandersetzung mit CDR in der Forschung statt: So fassen zum Beispiel Mihale-Wilson et al. (2022) und Müller (2022) den aktuellen Stand der CDR-bezogenen Forschung zusammen. CDR wurde sogar zum Hauptthema zentraler (deutscher) Konferenzen wie

der Wirtschaftsinformatik 2023 (vgl. Mirbabaie 2023) oder zum Untersuchungsgegenstand dedizierter Calls for Papers renommierter Journale (vgl. Recker et al. 2023; Schreck et al. 2021).

All diesen Entwicklungen zum Trotz gibt es weder in der Praxis noch in der wissenschaftlichen Diskussion eine einheitliche CDR-Definition, geschweige denn eine gemeinsame Definition beider Sphären. Vielmehr haben sich unterschiedliche Definitionen mit Überschneidungen und Gemeinsamkeiten herausgebildet; für einen Vergleich siehe z. B. Bednárová/Serpeninova (2023) oder Müller (2022). Zwei Dimensionen der Dynamik verdeutlichen die Schwierigkeit, eine gemeinsame Definition zu finden: Die ersten Definitionen (z. B. vom BVDW und der deutschen CDR-Initiative) setzten vor allem auf eine rechtliche Komponente und legten den Fokus darauf, dass CDR über gesetzliche Anforderungen hinaus gehen müsse. Dann änderte sich das regulatorische Umfeld seit 2015 grundlegend, insbesondere durch das EU-KI-Gesetz, das Gesetz über digitale Dienste (DSA) und das Gesetz über digitale Märkte (DMA). Gleichzeitig geschehen technologische Innovationen in einem enormen Tempo: Die Dynamik der Entwicklungszyklen sowie neue Innovationen, die auf den Markt kommen, werfen laufend neue Fragen zu ihren Auswirkungen und einem verantwortungsvollen Umgang mit ihnen auf.

Dieser Denkimpuls will die CDR-Debatte mit einer einheitlichen CDR-Definition voranbringen, ohne bestehende Definitionen zu verdrängen. Sie bleiben gültig und relevant, da sie in Teilbereichen mehr Tiefe bieten. Ein übergreifender, allgemeingültiger Konsens über eine hochrangige Definition von CDR kann jedoch den Diskurs vorantreiben und so der Debatte über CDR insgesamt mehr Kraft verleihen, um anschließend gezielt in die Details einzutauchen.

3. Methodisches Vorgehen und Entwicklung einer gemeinsamen Definition von Corporate Digital Responsibility

Der iterative Ansatz zur Entwicklung einer gemeinsamen CDR-Definition von Wissenschaft, Praxis, Zivilgesellschaft und öffentlichen Einrichtungen besteht aus drei Hauptphasen. Im ersten Schritt wurden bestehende Definitionen von CDR aus drei Bereichen geprüft: (i) Initiativen und Frameworks, die auf CDR abzielen, (ii) Unternehmen, die sich mit CDR beschäftigen, und (iii) wissenschaftliche Publikationen zu CDR. Die Vielfalt der unterschiedlichen CDR-Definitionen sowie der individuellen und organisatorischen Akteure, die in diesem Bereich tätig sind, macht

einen lückenlosen Überblick der im ersten Schritt untersuchten Definitionen unmöglich. Stattdessen werden im Folgenden nur einige der zentralsten CDR-Definitionen vorgestellt, die in Europa und im Vereinigten Königreich diskutiert werden und international Anwendung finden.

Initiative/ Framework	CDR-Definition	Quelle
CDR Award (BVDW)	„Corporate Digital Responsibility (CDR) bezeichnet digital verantwortliches Handeln über das gesetzlich Geforderte hinaus in Bezug auf alle Stufen der Wertschöpfungskette.“	BVDW (2024)
CDR Building Bloxx (BVDW)	„Corporate Digital Responsibility (CDR) befasst sich mit der Rolle unternehmerischer Verantwortung in einer zunehmend digitalisierten Welt. Die Reflexion über die Auswirkungen des eigenen Handelns auf die Gesellschaft und Umwelt ist ein essentieller Bestandteil einer ganzheitlich gedachten digitalen Transformation. CDR-Initiativen gehen dabei oft über bestehende regulatorische Anforderungen hinaus und erfordern proaktive Selbstverpflichtungsmaßnahmen.“	BVDW (2025)
CDR-Kodex (CDR-Initiative)	„Wir verstehen CDR dabei als freiwillige unternehmerische Aktivitäten, die insbesondere im Sinne der Verbraucherinnen und Verbraucher über das gesetzlich Vorgeschriebene hinausgehen und die digitale Welt aktiv zum Vorteil der Gesellschaft mitgestalten. CDR ist somit Teil einer umfassenden Unternehmensverantwortung.“	CDR-Initiative (2025: 2)
CDR Manifesto (Rob Price)	„Corporate Digital Responsibility (CDR) bündelt eine Reihe von Praktiken und Verhaltensweisen, die einem Unternehmen helfen, Daten und digitale Technologien auf eine Weise zu nutzen, die als sozial, wirtschaftlich und ökologisch verantwortungsvoll wahrgenommen wird.“	Price (2021)
Swiss Digital Initiative	„Corporate Digital Responsibility is a relatively new field that is gaining momentum as more aspects of our lives at work and beyond involve digital technologies. While digital technologies offer great benefits, they raise difficult ecological, social and ethical questions that need to be addressed. Corporate Digital Responsibility provides a framework to do just that and ensures that digital technologies are developed and used in a beneficial way for organisations, users and society at large.“	Swiss Digital Initiative (2025)
Zentrum Digitalisierung Bayern	„[…] die verantwortliche Gestaltung der Digitalisierung im Kerngeschäft des Unternehmens.“	Esselmann et al. (2020: 6)

ABBILDUNG 1: ÜBERBLICK: CDR-DEFINITIONEN VON INITIATIVEN / IN FRAMEWORKS (QUELLE: EIGENE ABBILDUNG)

Unternehmen	CDR-Definition	Quelle
Bayern Innovativ	„Corporate Digital Responsibility (CDR) [...]: die unternehmerische Verantwortung und das Nachhaltigkeitsmanagement im Digitalzeitalter.“	Bayern Innovativ (2025)
Deloitte	„Corporate Digital Responsibility (CDR) ergänzt die unternehmerische Verantwortung und denkt sie teilweise neu. Beispielsweise müssen Unternehmen die gesellschaftlichen Auswirkungen der digitalen Produkte und Dienstleistungen bereits in ihrer Entwicklung mitbedenken und sicherstellen, dass sie mit unseren Wertmaßstäben kompatibel sind.“	Deloitte (2023)
DIHK	„Corporate Digital Responsibility (abgekürzt: CDR) bedeutet, dass sich Unternehmen verpflichten, im Rahmen der Digitalisierung nachhaltig und verantwortungsvoll zu handeln.“	DIHK (2023)
Institut für ökologische Wirtschaftsforschung	„Das CDR-Konzept hat einen breiten Geltungsanspruch, das heißt es dient Unternehmen aller Branchen zur Orientierung, nicht nur IT-Unternehmen. Dabei verfolgt das Konzept einen integrativen Ansatz, indem es CDR in die etablierten Handlungsfelder der Unternehmensverantwortung einbettet.“	Lautermann/Frick (2023)
Pricewaterhouse-Coopers	„Corporate Digital Responsibility und Digitale Ethik sind ganzheitliche Konzepte und Perspektiven, die wir dazu nutzen, um gemeinsam mit unseren Kunden und unserem interdisziplinären Team die Digitalisierung vertrauenswürdig zu gestalten. Die beiden Konzepte sind eng miteinander verknüpft und bilden ein zentrales Element der integren Unternehmensführung. “	PricewaterhousCoopers (2025)
The Corporate Governance Institute	„Corporate digital responsibility (CDR) is a principle that demands businesses act as stewards of technology in a world increasingly shaped by digital power.“	Byrne (2025)
WiseWay	„Corporate Digital Responsibility (CDR) gehört als Bereich zu einer umfassenden Unternehmensverantwortung (CR) in einer zunehmend digitalisierten Wirtschaft und Gesellschaft. Es handelt sich um „freiwillige unternehmerische Aktivitäten im digitalen Bereich, die über das heute gesetzlich Vorgeschriebene hinausgehen und die digitale Welt aktiv zum Vorteil der Gesellschaft mitgestalten“ (BMJV 2018: 1). CDR bezieht sich einerseits auf die Beachtung digitaler Nachhaltigkeit (d.h. die Nachhaltigkeit von Daten und Algorithmen, vgl. Stürmer et al. 2017; Smart-Data-Begleitforschung 2018) und anderseits auf Berücksichtigung der sozialen, ökonomischen und ökologischen Wirkungen digitalen Unternehmenshandelns in der Welt (vgl. Esselmann/Brink 2016; Mühlner 2017; Thorun 2018).“	Dörr (2020)

ABBILDUNG 2: ÜBERBLICK: CDR-DEFINITIONEN VON UNTERNEHMEN (QUELLE: EIGENE ABBILDUNG)

Autor*innen	CDR-Definition	Quelle
Frank Esselmann, Alexander Brink	„[…] zwischen informationeller Selbstbestimmung (Digital Autonomy) und echter Wertschöpfung (Shared Value)."	Esselmann/ Brink (2016: 39)
K. Valerie Carl, Oliver Hinz	„The aim of CDR is to ensure a comprehensive understanding on how to use the opportunities as well as address the related risks of digitalization, in this way contributing to human and social value creation and fostering behavior that is aligned with ethical norms and values. Such a comprehensive concept better reflects reality where responsibilities do not occur isolated but rather in an interconnected way with potential influences on one another."	Carl/Hinz (2024: 2)
Christina J. Herden et al.	„[W]e regard CDR as an extension of CSR, comprising all levels of corporate responsibilities as defined in Carroll's (1991) CSR pyramid and all domains of the Environmental, Social, Governance (ESG) framework."	Herden et al. (2021: 14)
Lara Lobschat et al.	„We refer to this idea as corporate digital responsibility (CDR), defined as the set of shared values and norms guiding an organization's operations with respect to the creation and operation of digital technology and data."	Lobschat et al. (2021: 876)
Cristina Mihale-Wilson et al.	„CDR intends to minimize the adverse effects of digitalization while maximizing the positive impacts of corporate digital activities. In this vein, CDR seeks to ensure an ethical and responsible development, deployment, and use of digital technologies and data."	Mihale-Wilson et al. (2021: 1)
Cristina Mihale-Wilson et al.	„CDR articulates companies' extended responsibilities regarding the new opportunities and challenges that technology development and use can bring (Herden et al. 2021). At its core, CDR is related to Corporate Social Responsibility (CSR). In essence, both concepts summarize the obligations that companies have toward society. Moreover, both concepts aim at minimizing the adverse effects of practice activities while maximizing the benefits of a firm's activities."	Mihale-Wilson et al. (2022 : 127)
Benjamin Müller	„CDR is concerned with ensuring that corporations exhibit behaviors that comply with a larger understanding of good or positive behaviors. […] CDR serves an important governance function in that a CDR regime also seeks to define how to effectuate corporate behaviors across levels that are compliant with the norms and values […]."	Müller (2022: 692 f.)
Jochen Wirtz et al.	„We define CDR in the context of service as the principles underpinning a service firm's ethical, fair, and protective use of data and technology when engaging with customers within their digital service ecosystem."	Wirtz et al. (2023: 1)

ABBILDUNG 3: ÜBERBLICK: CDR-DEFINITIONEN IN DER FORSCHUNG (QUELLE: EIGENE ABBILDUNG)

Im zweiten Schritt diskutierten die beiden im Definitionsprozess federführenden Expert*innen, gestützt auf die überprüften Definitionen aus Praxis und Theorie, einen Vorschlag für eine ge-

meinsame Definition von CDR. Ziel war eine prägnante Ein-Satz-Formulierung, um ihre Anwendbarkeit in der Praxis, in der Lehre, in Expert*innen- und Nicht-Expert*innenforen und in der Forschung gleichermaßen zu gewährleisten. Dafür holten sie schriftliches Feedback von sechs weiteren CDR-Expert*innen ein, um die Definition des Konzepts zu schärfen. Bei der Entwicklung einer neuen, übergreifenden Definition von CDR wurden veränderte externe Effekte (z. B. Regulierung) berücksichtigt. Das Ergebnis, also die erste Version der High-Level-Definition des CDR-Konzepts, lautet wie folgt:

> CDR bietet einen ganzheitlichen Rahmen, um aktiv über die gesamte Wertschöpfungskette hinweg mit den gestiegenen Anforderungen für verantwortungsbewusstes Handeln von Unternehmen im digitalen Zeitalter umzugehen.

Die beiden federführenden Expert*innen schärften die Arbeitsdefinition in mehreren Iterationsschleifen. Schriftliches Feedback von sechs weiteren CDR-Expert*innen floss ein, was zu einer Anpassung einiger Aspekte und des Wortlauts der vorläufigen Definition führte. Dieser iterative Feedback-Zyklus führte zur folgenden, leicht angepassten vorläufigen Arbeitsdefinition von CDR:

> CDR bietet einen ganzheitlichen Rahmen sowie Maßnahmen, um aktiv – auch über die Wertschöpfungskette hinweg – mit den Anforderungen für ein verantwortungsbewusstes Handeln von Unternehmen in einer zunehmend digitalisierten Welt umzugehen.

Im dritten Schritt wurde ein Workshop mit 15 CDR-Expert*innen abgehalten, um die endgültige Definition zu entwickeln und einen breiten Konsens darüber zu erreichen. Die meisten dieser Expert*innen nahmen mehrere Perspektiven ein, aber primär sind vier der Wissenschaft, sechs der Praxis und fünf der Zivilgesellschaft und öffentlichen Einrichtungen zuzuordnen. Der Workshop bestand aus zwei Hauptteilen, nämlich (i) der Bewertung der Arbeitsdefinition und anschließend (ii) der Priorisierung. Um die vorläufige CDR-Arbeitsdefinition gemeinsam zu verfeinern, arbeiteten die 15 CDR-Expert*innen mit einem Mural Board. Sie bewerteten die Arbeitsdefinition zunächst einzeln und unabhängig voneinander: Auf einer fünfstufigen Skala markierten sie ihre

Zustimmung zu bestimmten Aspekten (z. B. Perspektiven, Formulierungen, Ansätze) der vorläufigen CDR-Definition. Dieser Ansatz machte Gemeinsamkeiten und Unterschiede in der Wahrnehmung der Definition und damit notwendige Anpassungen sichtbar. Anschließend diskutierten sie diese Aspekte. Bei der Auswertung der geposteten Meinungen und der anschließenden Diskussion stachen acht Hauptpunkte hervor:

1. Der Begriff „Wertschöpfungskette" wurde sehr unterschiedlich wahrgenommen und führte zu intensiven Diskussionen. Er verdeutlichte zwar den Unternehmensfokus der Definition, weckte aber starke Assoziationen zur Herstellung und passt daher nicht zum Ziel einer branchenübergreifend anwendbaren Definition.
2. Das Wort „aktiv" allein greift zu kurz: Es vermittelt eher ein durch äußere Umstände ausgelöstes Verhalten, statt die echte Eigeninitiative von Unternehmen zu betonen. Die Expert*innen waren sich weitgehend einig, dass dies angepasst werden sollte.
3. In diesem Sinne wurden auch die „Anforderungen" als zu passiv empfunden. Die Expert*innen plädierten für eine Formulierung, die den gestaltenden Charakter von CDR stärker betont.
4. Der Begriff „digitalisierte Welt" war einigen Expert*innen, insbesondere denen aus der wissenschaftlichen Debatte, zu unspezifisch. Man einigte sich jedoch darauf, sich nicht auf bestimmte Technologien zu konzentrieren, um die Definition langfristig anschlussfähig zu halten.
5. Um den Begriff „verantwortliches Handeln" klarer zu fassen, schlugen einige Teilnehmer*innen vor, das Wesen der Verantwortung durch „Dimensionen der Verantwortung" zu spezifizieren, die ausdrücklich die Schlüsseldimensionen der Verantwortung – die gesellschaftliche, die ökologische und die ökonomische – widerspiegeln.
6. Die Teilnehmenden diskutierten, dass CDR nicht nur das interne Unternehmensverhalten, sondern auch externe Auswirkungen digitaler Unternehmenstätigkeiten innerhalb „der Ökosysteme der Interessengruppen" widerspiegeln soll – von unbeabsichtigten Folgen (z. B. Verzerrungen in KI-Systemen, digitale Kluft) bis zu positiven Auswirkungen (z. B. digitale Barrierefreiheit, grüne IT). Die Anerkennung dieser „systemischen Perspektive" trägt dazu bei, CDR als proaktiven Governance-Ansatz von reaktivem Risikomanagement zu unterscheiden.
7. Die Expert*innen sprachen sich vor allem für eine „einfache, verständliche Definition" aus, die nicht zu komplex oder lang ist und keine weiteren Erklärungen erfordert. Praxis wie Forschung sollten die Definition sofort verstehen können.

8. Ein Spannungsfeld ergab sich bei der Frage, wie gut die Definition in unterschiedlichen Kontexten – etwa Bildung, wissenschaftliche Veröffentlichungen oder Praxispräsentationen – anwendbar ist.

Im zweiten Teil des Workshops bauten die Expert*innen auf diesen Diskussionen auf. Sie priorisierten die diskutierten Aspekte zunächst einzeln nach ihrer Relevanz für die Konsensfindung und konnten bei Bedarf neue Aspekte hinzuzufügen. Anschließend diskutierten sie die dringlichsten Aspekte im Plenum und kamen jeweils zu einer Einigung. Auf der Grundlage dieser Diskussionen entwickelten sie die Definition in zehn Iterationsschleifen weiter, bis ein Konsens – das vorher vereinbarte Abbruchkriterium – erreicht wurde. Die acht Kernpunkte aus dem ersten Teil wurden dabei wie folgt behandelt:

1. Auf den Begriff „Wertschöpfungskette" wurde verzichtet, da das dahinterliegende Anliegen durch den Begriff „ganzheitlich" und das gemeinsame Verständnis davon bereits erfasst wird.
2. Die beiden Hauptkritiken – zu passive Formulierung und unklare Verantwortungsebene – wurden adressiert, indem die „Handlungen" der Unternehmen herausgestellt und der spezifische Aspekt der „Auswirkungen" von CDR integriert wurde, wobei sowohl unbeabsichtigte Folgen als auch positive Effekte erfasst werden. CDR sollte ein eigenständiges Verhalten oder eine eigenständige Handlungsweise beinhalten und nicht nur ein implizites Verhalten sein.
3. Die Teilnehmenden einigten sich darauf, „digitalisierte Welt" durch „digitale Technologien und Daten" zu ersetzen, um ein Gleichgewicht zwischen der langfristigen Anwendbarkeit der Definition und dem Bezug zur Praxis herzustellen.
4. Im Sinne einer kürzeren und weniger komplexen Definition verzichteten die Expert*innen schließlich darauf, detailliert darzulegen, welche „Dimensionen von Verantwortung" durch CDR abgedeckt werden. Dennoch ist es sowohl in der Praxis als auch in der Theorie wichtig, sich der verschiedenen Facetten von Verantwortung bewusst zu sein.

Insgesamt führten die Diskussionen zu einer wesentlichen Kürzung der Definition und eine aktivere Formulierung zu einem weniger komplexen Satz. Die Expert*innen aus den verschiedenen Interessengruppen einigten sich auf die entwickelte Definition als übergeordnete Definition, die in verschiedenen Kontexten anwendbar ist. Zugleich erkannten sie an, dass die Forschung und die Praxis die Definition bei der Verwendung in unterschiedliche Kontexte stellen werden. Dieser iterative Prozess führte somit zur folgenden, vereinbarten Definition von CDR (in Englisch und Deutsch):

CDR stands for holistic responsible corporate action with regard to the impact of digital technologies and data.

CDR steht für ganzheitliches verantwortungsbewusstes Handeln von Unternehmen in Bezug auf die Auswirkungen digitaler Technologien und Daten.

4. Definition als Nährboden für die Weiterentwicklung von Corporate Digital Responsibility

Ziel dieses Denkimpulses war es, iterativ eine gemeinsame Definition von CDR zu entwickeln, die von verschiedenen Interessengruppen (Praxis, Zivilgesellschaft, öffentliche Einrichtungen, Forschende unterstützt wird und die auch die Gemeinsamkeiten der bestehenden Definitionen von CDR bündelt. Die neue Formulierung soll das Konzept von CDR wissenschaftlich und praktisch voranbringen, ohne die derzeit diskutierten Definitionen von CDR zu ersetzen. Vielmehr macht sie sichtbar, worüber sich die verschiedenen Stakeholder*innen im Bereich CDR bereits einig sind.

Dafür wurde das gemeinsame Verständnis des Begriffs CDR aus verschiedenen Perspektiven zusammengefasst. Allerdings erfordern diese unterschiedlichen Perspektiven auch unterschiedliche Akzente, die diese Definition in Zukunft ergänzen könnten, z. B. durch zusätzliche Informationen wie die gesellschaftliche, ökologische und wirtschaftliche Dimension, um die Definition für die Praxis greifbarer zu machen. Auch wenn die endgültige Definition bewusst knappgehalten ist, trägt die Anerkennung dieser Wirkungsbereiche zu einem klareren Verständnis dessen bei, wie CDR über die allgemeine Unternehmensverantwortung hinausgeht und mit Nachhaltigkeitszielen in Einklang steht.

Die entwickelte Definition unterstützt die Operationalisierung von CDR, indem sie ein allgemein anerkanntes Verständnis von CDR schafft und damit besonders Nicht-Expert*innen den praktischen Zugang zum Konzept erleichtert. Die Definition soll einen wichtigen Eckpfeiler für eine zunehmende Operationalisierung in der Praxis sowie deren empirische Erforschung bilden. Die entwickelte Definition sollte eine lebendige Definition sein, die im Laufe der Zeit angepasst werden kann. Trotzdem soll die gemeinsame CDR-Definition die Debatte vorantreiben und zeigen, dass CDR den Kinderschuhen entwachsen ist.

Adjerid, I. / Peer, E. / Acquisti, A. (2018): Beyond the Privacy Paradox: Objective Versus Relative Risk in Privacy Decision Making, in: MIS Quarterly, Jg. 42 / Nr. 2, 465–488.

Bayern Innovativ (2025): Corporate Digital Responsibility – effizient, erfolgreich und klimabewusst, URL: https://www.bayern-innovativ.de/emagazin/transformation/detail/corporate-digital-responsibility-effizient-erfolgreich-und-klimabewusst/ (aufgerufen am: 14/08/2025).

Bednárová, M. / Serpeninova, Y. (2023): Corporate Digital Responsibility: Bibliometric Landscape – Chronological Literature Review, in: The International Journal of Digital Accounting Research, Jg. 23 / Nr. 29, 1–18.

Bélanger, F. / Crossler, R. E. (2011): Privacy in the Digital Age: A Review of Information Privacy Research in Information Systems, in: MIS Quarterly, Jg. 35 / Nr. 4, 1017–1041.

Benbya, H. / Pachidi, S. / Jarvenpaa, S. L. (2021): Special Issue Editorial: Artificial Intelligence in Organizations: Implications for Information Systems Research, in: Journal of the Association for Information Systems, Jg. 22 / Nr. 2, 281–303.

BMJV (2018): Corporate Digital Responsibility-Initiative: Digitalisierung verantwortungsvoll gestalten. Eine gemeinsame Plattform, URL: https://www.bmjv.de/DE/themen/verbraucherschutz/digitaler_verbraucherschutz/cdr/cdr.html (aufgerufen am: 14/08/2025).

Brîncoveanu, C. / Carl, K. V. / Binz, S. / Weiher, M.-A. / Thomas, O. / Hinz, O. (2024): Artificial Intelligence-Based Assistance Systems for Environmental Sustainability in Smart Homes: A Systematic Literature Review on Requirements and Future Directions, in: Klein, M. / Krupka, D. / Winter, C. / Gergeleit, M. / Martin, L. (Hrsg.): INFORMATIK 2024. Lecture Notes in Informatics (LNI), Bonn: Gesellschaft für Informatik, 1165–1182.

BVDW (2024): CDR Award, URL: https://www.cdr-award.digital/ (aufgerufen am: 14/08/2025).

– (2025): CDR Building Bloxx, URL: https://www.cdr-building-bloxx.com/ (aufgerufen am: 14/08/2025).

Byrne, D. (2025): What is Corporate Digital Responsibility?, URL: https://www.thecorporategovernanceinstitute.com/insights/lexicon/what-is-corporate-digital-responsibility/ (aufgerufen am: 14/08/2025).

CDR-Initiative (2025): Corporate Digital Responsibility-Kodex. Freiwillige Selbstverpflichtung mit Bericht, URL: https://cdr-initiative.de/uploads/files/2024-01_Kodex_CDR-Initiative.pdf (aufgerufen am: 14/08/2025).

Carl, K. V. / 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: Integrating Understandings, Definitions, and Expert Perspectives, Creative Commons BY-SA 4.0, in: Informatik 2025 [im Erscheinen].

Carl, K. V. / Hinz, O. (2024): What We Already Know About Corporate Digital Responsibility in IS Research: A Review and Conceptualization of Potential CDR Activities, in: Electronic Markets, Jg. 34 / Nr. 1, Article 27, URL: https://link.springer.com/article/10.1007/s12525-024-00708-0 (aufgerufen am: 14/08/2025).

Carl, K. V. / Kubach, M. / Mihale-Wilson, C. (2023): The Motivation of Companies to Implement Corporate Digital Responsibility Activities Voluntarily: An Empirical Assessment, in: Beverungen, D. / Lehrer, C. / Trier, M. (Hrsg.): Conceptualizing Digital Responsibility for the Information Age. Proceedings of the 18th International Conference on Wirtschaftsinformatik, Paderborn, 39–52.

Cooper, T. / Siu, J. / Wei, K. (2015): Corporate Digital Responsibility. Doing Well by Doing Good. Accenture, URL: https://criticaleye.com/inspiring/insights-servfile.cfm?id=4431 (aufgerufen am: 14/08/2025).

Deloitte (2023): Corporate Digital Responsibility, URL: https://www.deloitte.com/de/de/services/consulting/research/corporate-digital-responsibility.html (aufgerufen am: 14/08/2025).

DIHK (2023): Corporate Digital Responsibility – CDR Definition, URL: https://ihk-kompetenz.plus/wissenswertes/corporate-digital-responsibility-cdr-definition/ (aufgerufen am: 14/08/2025).

Dörr, S. (2022): Begriffsbestimmung Corporate Digital Responsibility, URL: https://wiseway.de/project/begriffsbestimmung-corporate-digital-responsibility/ (aufgerufen am: 14/08/2025).

Esselmann, F. / Brink, A. (2016): Corporate Digital Responsibility. Den digitalen Wandel von Unternehmen und Gesellschaft erfolgreich gestalten, in: Spektrum. Das Wissenschaftsmagazin der Universität Bayreuth, Jg. 12 / Nr. 1, 38–41, URL: https://www.presse.uni-bayreuth.de/de/spektrum-archiv/_pdf/ausgabe_01_16.pdf (aufgerufen am: 14/08/2025).

Esselmann, F. / Golle, D. / Thiel, D. C. / Brink, A. (2020): Corporate Digital Responsibility: Unternehmerische Verantwortung als Chance für die deutsche Wirtschaft. ZD.B Whitepaper, Bayern, München: Zentrum Digitalisierung Bayern, URL: https://cdr.bayern.de/cdr_wissen/publikationen/doc/ZD.B%20Positionspapier_Corporate%20Digital%20Responsibility.pdf (aufgerufen am: 14/08/2025).

Herden, C. J. / Alliu, E. / Cakici, A. / Cormier, T. / Deguelle, C. / Gambhir, S. / Griffiths, C. / Gupta, S. / Kamani, S. R. / Kiratli, Y.-S. / Kispataki, M. / Lange, G. / Moles de Matos, L. / Tripero Moreno, L. / Beancourt Nunez, H. A. / Pilla, V. / Raj, B. / Roe, J. / Skoda, M. / Song, Y. / Ummadi, P. K. / Edinger-Schons, L. M. (2021): Corporate Digital Responsibility: New Corporate Responsibilities in the Digital Age, in: Sustainability Management Forum, Jg. 29 / Nr. 2, 13–29, URL: https://link.springer.com/article/10.1007/s00550-020-00509-x (aufgerufen am: 14/08/2025).

Keller, K. (2020): Intelligent Parking Systems – The Smart Assistant Thinks Around the Corner, in: George, J. F. / Paul, S. / De', R. / Karahanna, E. / Sarker, S. / Oestereicher-Singer, G. (Hrsg.): Proceedings of the 41st International Conference on Information Systems (ICIS 2020), Hyderabad: Association for Information Systems.

Kordzadeh, N. / Ghasemaghaei, M. (2022): Algorithmic Bias: Review, Synthesis, and Future Research Directions, in: European Journal of Information Systems, Jg. 31 /Nr. 3, 388–409.

Lautermann, C. / Frick, V. (2023): Corporate Digital Responsibility: Wie Unternehmen im digitalen Wandel Verantwortung übernehmen, Schriftenreihe des IÖW, Berlin: Institut für ökologische Wirtschaftsforschung, URL: https://www.ioew.de/fileadmin/user_upload/BILDER_und_Downloaddateien/Publikationen/2022/Lautermann_Frick2022-CDR-Grundlagenstudie.pdf (aufgerufen am: 14/08/2025).

Lobschat, L. / Müller, B. / Eggers, F. / Brandimarte, L. / Diefenbach, S. / Kroschke, M. / Wirtz, J. (2021): Corporate Digital Responsibility, in: Journal of Business Research, Jg. 122 / Nr. 1, 875–888, URL: https://www.sciencedirect.com/science/article/pii/S0148296319305946 (aufgerufen am: 14/08/2025).

Mädche, A. / Legner, C. / Benlian, A. / Berger, B. / Gimpel, H. / Hess, T. / Hinz, O. / Morana, S. / Söllner, M. (2019): AI-Based Digital Assistants, in: Business & Information Systems Engineering, Jg. 61 / Nr. 4, 535–544.

Mihale-Wilson, C. / Hinz, O. / van der Aalst, W. / Weinhardt, C. (2022): Corporate Digital Responsibility: Relevance and Opportunities for Business and Information Systems Engineering, in: Business & Information Systems Engineering, Jg. 64 / Nr. 2, 127–132, URL: https://link.springer.com/article/10.1007/s12599-022-00746-y (aufgerufen am: 14/08/2025).

Mihale-Wilson, C. / Zibuschka, J. / Carl, K. V. / Hinz, O. (2021): Corporate Digital Responsibility – Extended Conceptualization and a Guide to Implementation, in: ECIS 2021 Research Papers 80.

Mikalef, P. / Conboy, K. / Lundström, J. E. / Popovič, A. 2022 : Thinking Responsibly About Responsible AI and "the Dark Side" of AI, in: European Journal of

Information Systems, Jg. 31 / Nr. 3, 257–268, URL: https://www.tandfonline.com/doi/epdf/10.1080/0960085X.2022.2026621?needAccess=true (aufgerufen am: 14/08/2025).

Mirbabaie, M. (2023): WI2023 Konferenz, URL: http://wi2023.de/ (aufgerufen am: 14/08/2025).

Müller, B. (2022): Corporate Digital Responsibility, in: Business & Information Systems Engineering, Jg. 64 / Nr. 5, 689–700, URL: https://link.springer.com/article/10.1007/s12599-022-00760-0 (aufgerufen am: 14/08/2025).

Mühlner, J. (2017): Corporate Digital Responsibility: Verantwortung in der digitalen Gesellschaft, in: Forum Europrofession.

Price, R. (2021): CDR Manifesto. Die sieben Grundprinzipien der internationalen Definition der digitalen Verantwortung von Unternehmen, URL: https://corporatedigitalresponsibility.net/cdr-definition-german (aufgerufen am: 14/08/2025).

PricewaterhouseCoopers (2025): Corporate Digital Responsibility und Digitale Ethik, URL: https://www.pwc.de/de/nachhaltigkeit/corporate-digital-responsibility-und-digitale-ethik.html (aufgerufen am: 14/08/2025).

Ransbotham, S. / Fichman, R. G. / Gopal, R. / Gupta, A. (2016): Special Section Introduction – Ubiquitous IT and Digital Vulnerabilities, in: Information Systems Research, Jg. 27 / Nr. 4, 834–847.

Recker, J. / Chatterjee, S. / Sundermeier, J. / Tarafdar, M. (2023): JAIS Special Issue Call for Participation: Digital Responsibility: Social, Ethical, and Ecological Implications of IS, in: Journal of the Association for Information Systems, URL: https://aisel.aisnet.org/jais/SI-DigitalResponsibility.pdf (aufgerufen am: 14/08/2025).

Santos, J. / Rodrigues, J. J. P. C. / Casal, J. / Saleem, K. / Denisov, V. (2016): Intelligent Personal Assistants Based on Internet of Things Approaches, in: IEEE Systems Journal, Jg. 12 / Nr. 2, 1793–1802.

Schreck, P. / Edinger-Schons, L.-M. / Uhl, M. (2021): Editorial: The Ethics of Digitalization and Emerging Corporate Responsibilities in the Digital Age, in: Journal of Business Economics, Jg. 94 / Nr. 7 975–978, URL: https://link.springer.com/article/10.1007/s11573-024-01202-x (aufgerufen am: 14/08/2025).

Smart-Data-Begleitforschung (2018): Corporate Digital Responsibility, Berlin: Fachgruppe Wirtschaftliche Potenziale & gesellschaftliche Akzeptanz.

Stürmer, M. / Abu-Tayeh, G. / Myrach, T. (2017): Digital Sustainability: Basic Conditions for Sustainable Digital Artifacts and Their Ecosystems, .in: Sustainability Science, Jg. 12 / Nr. 2, 247–262, URL: https://link.springer.com/article/10.1007/s11625-016-0412-2 (aufgerufen am: 14/08/2025).

Swiss Digital Initiative (2025): Corporate Digital Responsibility, URL: http://swiss-digital-initiative.org/corporate-digital-responsibility (aufgerufen am: 14/08/2025).

Tarafdar, M. / Gupta, A. / Turel, O. (2013): The Dark Side of Information Technology Use, in: Information Systems Journal, Jg. 23 / Nr. 3, 269–275.

Thorun, C. (2018): Corporate Digital Responsibility: Unternehmerische Verantwortung in der digitalen Welt, in: Gärtner, C. / Heinrich, C. (Hrsg.): Fallstudien zur Digitalen Transformation, Wiesbaden: Springer Gabler.

Vial, G. (2019): Understanding Digital Transformation: A Review and a Research Agenda, in: The Journal of Strategic Information Systems, Jg. 28 / Nr. 2, 118–144.

Wirtz, J. / Kunz, W. H. / Hartley, N. / Tarbit, J. (2023): Corporate Digital Responsibility in Service Firms and Their Ecosystems, in: Journal of Service Research, Jg. 26 / Nr. 2, 173–190, URL: https://journals.sagepub.com/doi/epub/10.1177/10946705221130467 (aufgerufen am: 14/08/2025).

Wong, R. Y. M. / Cheung, C. M. K. / Xiao, B. / Thatcher, J. B. (2021): Standing Up or Standing By: Understanding Bystanders' Proactive Reporting Responses to Social Media Harassment, in: Information Systems Research, Jg. 32 / Nr. 2, 561–581, URL: https://pubsonline.informs.org/doi/epdf/10.1287/isre.2020.0983 (aufgerufen am: 14/08/2025).

Das CDR-Reifegradmodell[1]

Rüstzeug für digitale Verantwortung in turbulenten Zeiten

Alexander Brink, Frank Esselmann und Leonhard Henke

1. Die Herausforderung der Zwillingstransformation

Die Twin Transformation beschreibt den gleichzeitigen Wandel hin zu Digitalisierung und Nachhaltigkeit – ein Konzept, das 2021 durch Accenture geprägt wurde (vgl. Accenture 2021). Unternehmen, die in beiden Bereichen stark aufgestellt sind – sogenannte Twin Transformers – gelten als besonders widerstands- und zukunftsfähig. Dies liegt daran, dass sie digitale Technologien gezielt für nachhaltige Geschäftsmodelle nutzen, Ressourcen strategisch bündeln und Verantwortung auch jenseits finanzieller Ziele übernehmen. Zudem fördern sie Partnerschaften für nachhaltige Wertschöpfung und beziehen Mitarbeitende aktiv in den Transformationsprozess ein (vgl. ebd.). Studien zeigen: Unternehmen mit integrierten Digital- und Nachhaltigkeitsstrategien erzielen deutlich bessere Ergebnisse als solche mit einseitigem Fokus. Eine kombinierte Herangehensweise erweist sich dabei als besonders wirksam. Dennoch bleibt der Diskurs oft oberflächlich: Während ökonomische Chancen betont werden, bergen digitale Innovationen Risiken, etwa durch Rebound-Effekte, die Effizienzgewinne neutralisieren. Die nachhaltige Transformation braucht daher eine normative Orientierung.

---

1 Dieser Beitrag wird in einer längeren englischsprachigen Version veröffentlicht in: The CDR Maturity Model. Guiding Digital Responsibility in Turbulent Times, in: Christ, A.-S. (Hrsg.) (2025): Digital Responsibility: From Principles to Practice, Berlin: Springer [im Erscheinen]. Das generative KI-Tool (OpenAI) wurde in unterstützender Funktion für stilistische Verfeinerungen, sprachliche Optimierungen und als Inspiration für alternative Formulierungen eingesetzt. In wenigen Fällen wurden erste Entwürfe von Passagen mit dem Tool erstellt und anschließend von den Autoren grundlegend überarbeitet und umformuliert. Alle intellektuellen Inhalte, kritischen Überlegungen und endgültigen Entscheidungen liegen in der alleinigen Verantwortung der Autoren.

2. *Die Corporate-Digital-Responsibility-Reise*

2.1 CDR als Schlüssel zur menschenzentrierten Digitalisierung

Die digitale Transformation verändert Wirtschaft und Gesellschaft grundlegend. Neue Technologien wie Automatisierung, Datenanalyse und vernetzte Systeme bieten große Chancen. Gleichzeitig wirft dieser Wandel zentrale Fragen auf: Wie kann Technik im Sinne der Menschen gestaltet werden, nicht bloß zugunsten wirtschaftlicher Interessen? Wie lassen sich Daten verantwortungsvoll einsetzen und digitale Selbstbestimmung sichern? Und wie gelingt eine faire, inklusive und nachhaltige Digitalisierung?

Hier setzt die Corporate Digital Responsibility (CDR) an. Sie beschreibt die Verantwortung von Unternehmen, digitale Technologien sowie Daten sozial, ökologisch und wirtschaftlich verantwortungsvoll zu nutzen. CDR geht dabei über die Erfüllung gesetzlicher Vorschriften hinaus: Sie will aktiv zum Gemeinwohl und zu einer gerechten, resilienten Gesellschaft beitragen. Im Zentrum steht eine menschenzentrierte Digitalisierung, die sich an Werten, Rechten und Bedürfnissen orientiert (vgl. Brink 2025). CDR umfasst zentrale Themen wie Umgang mit Daten, Bildung, Klima- und Ressourcenschutz, Mitarbeitenden-Einbindung und Inklusion (vgl. CDR-Initiative 2021). Diese fünf Handlungsfelder des CDR-Kodex geben Unternehmen Orientierung für eine wertebasierte Gestaltung digitaler Innovationen. Ziel ist es, Vertrauen zu stärken und Werte wie Selbstbestimmung, Gerechtigkeit und Nicht-Diskriminierung zu sichern. So betont Grundsatz 5 „Autonomie", dass Nutzer selbst über ihre Daten entscheiden sollen. Grundsatz 1 fordert, dass digitale Systeme demokratische Grundwerte wie Gleichbehandlung und Freiheit wahren (vgl. ebd.). Wissenschaftlich versteht man CDR als ein „set of shared values and norms guiding an organization's operations with respect to the creation and operation of digital technology and data" (Lobschat et al. 2021: 876). CDR ist eng mit digitaler Ethik verknüpft, die sich mit moralischen Fragen der Technologie-Nutzung befasst (vgl. Brink 2022).

Zugleich zeigt sich die Digitalisierung ambivalent: Sie kann medizinische Diagnosen verbessern und Leben retten, aber auch Diskriminierung fördern – etwa durch voreingenommene KI bei Bewerbungen oder Krediten. Diese Gegensätze verdeutlichen, dass Digitalisierung aktiv und menschenorientiert gestaltet werden muss (vgl. Brink 2025). CDR bietet dazu den ethischen und strategischen Rahmen, um Chancen verantwortungsvoll zu nutzen und Risiken zu minimieren.

2.2 Die Rolle der CDR-Initiative und die Formen der Zusammenarbeit

Die im Jahr 2018 vom Bundesministerium der Justiz und für Verbraucherschutz ins Leben gerufene CDR-Initiative beschreibt Corporate Digital Responsibility (CDR) als:

> [...] freiwillige unternehmerische Aktivitäten, die insbesondere im Sinne der Verbraucherinnen und Verbraucher über das gesetzlich Vorgeschriebene hinausgehen und die digitale Welt aktiv zum Vorteil der Gesellschaft mitgestalten (CDR-Initiative 2021).

Die CDR-Initiative fördert die strategische Verankerung digitaler Verantwortung in Unternehmen. Sie unterstützt bei der werteorientierten Weiterentwicklung von Prozessen und Produkten – methodisch und inhaltlich, etwa durch Workshops, Leitfäden und Dialogformate. International wird sie als eine der fortschrittlicheren Initiativen wahrgenommen.

> The launch of the German CDR-code in June 2021 is a milestone in capturing the threats from digital transformation through voluntary commitments from signatories […] (Merbecks 2023: 1017).

Um Unternehmen und Organisationen in unterschiedlichen Entwicklungsstadien und mit unterschiedlichen Zielen die Teilnahme zu ermöglichen, bietet die CDR-Initiative drei klar differenzierte Formen der Zusammenarbeit an. Diese Modelle ermöglichen einen flexiblen Zugang zur Initiative – ob als Impulsgeber, Lernpartner oder Umsetzungstreiber.

Mitgliedschaft

Die Mitgliedschaft richtet sich an Unternehmen, die einen langfristigen und strategischen Beitrag zur Weiterentwicklung der digitalen Verantwortung von Unternehmen leisten wollen. Mitglieder sind aktiv in das Netzwerk der CDR-Initiative eingebunden, bringen eigene Erfahrungen ein und profitieren gleichzeitig vom Austausch mit anderen Organisationen, Wissenschaft, Politik und Zivilgesellschaft. Sie tragen zu zentralen Themen bei, gestalten die Inhalte der Initiative mit und unterstützen die Verbreitung der CDR-Prinzipien in Wirtschaft und Gesellschaft.

Themenpartnerschaft

Themenpartnerschaften bieten die Möglichkeit einer gezielten Zusammenarbeit zu bestimmten Themen oder Projekten. Diese Form der Zusammenarbeit eignet sich besonders für Organisationen, die ihr Fachwissen oder ihre Interessen in bestimmten Bereichen wie Datenethik, Vielfalt, algorithmische Fairness oder Desinformation einbringen möchten. Themenpartnerschaften beteiligen sich punktuell, z. B. in Arbeitsgruppen, Studien, Pilotprojekten oder Dialogformaten – und leisten so einen wichtigen Beitrag zur thematischen Vertiefung und praktischen Umsetzung des CDR.

Entwicklungspartnerschaft

Die Entwicklungspartnerschaft richtet sich an Organisationen, die Corporate Digital Responsibility systematisch in ihre Strukturen und Prozesse integrieren wollen. Sie erhalten von der CDR-Initiative methodische und inhaltliche Unterstützung – z. B. durch Workshops, Coaching, Reifegradanalysen und Aktionsplanung. Ziel ist es, CDR unternehmensspezifisch zu verankern, die interne Verankerung zu fördern und nachhaltige Entwicklungsziele im digitalen Kontext zu erreichen. Die Entwicklungspartner profitieren von erprobten Instrumenten und dem Zugang zu Erfahrungswerten aus dem CDR-Netzwerk.

2.3 Die Neuausrichtung der CDR-Initiative: Flexibel, fokussiert, wirkungsorientiert

Die Relevanz von Corporate Digital Responsibility wächst – und damit auch die Anforderungen an die Initiative selbst. Wir haben in den letzten Jahren hart und beharrlich daran gearbeitet, dies voranzutreiben, aber wir lernen auch noch und entwickeln uns weiter. Denn CDR kann nicht isoliert betrachtet werden, sondern steht in direktem Zusammenhang mit aktuellen gesellschaftlichen, politischen und wirtschaftlichen Entwicklungen – die aktuellsten und dringendsten davon sind:

- Technologischer Wandel: KI, Quantencomputing und Neurotechnologien verändern Wirtschaft und Alltag grundlegend. Digitale und reale Welt verschmelzen zunehmend.
- Sozialer Druck: Digitale Teilhabe ist ungleich verteilt. Während einige vom Fortschritt profitieren, fühlen sich andere ausgeschlossen oder überfordert.
- Geopolitische Abhängigkeiten: Europa bleibt bei digitalen Schlüsseltechnologien oft abhängig von anderen Staaten. Souveränität, Sicherheit und Ethik werden zur strategischen Aufgabe.

- Wirtschaftliche Transformation: Marktverzerrungen durch Tech-Monopole treffen aufsteigende Nachhaltigkeitsanforderungen. Ökologie und Innovation müssen zusammengedacht werden.

Die rasante digitale und gesellschaftliche Entwicklung stellt auch die CDR-Initiative vor neue Anforderungen. Ihre Bedeutung wächst – insbesondere mit Blick auf ethische Fragen, Verantwortung im digitalen Raum und soziale Auswirkungen. Gleichzeitig kann sie nicht alle Themen gleichzeitig abdecken: Die digitale Transformation ist vielschichtig, schnell und berührt technologische, regulatorische und gesellschaftliche Bereiche. CDR darf dabei nicht nur als Teil der Nachhaltigkeitsdebatte verstanden werden. Zwar bestehen inhaltliche Schnittmengen – etwa bei Teilhabe, Lieferketten oder ökologischer Verantwortung –, doch CDR bringt eigene Themen mit: etwa Datenethik, algorithmische Transparenz oder digitale Autonomie.

Damit CDR wirksam bleibt, braucht es klare Schwerpunkte und Offenheit für neue Entwicklungen. Drei zentrale Erkenntnisse leiten die Weiterentwicklung:

1. Branchenübergreifendes Lernen ist entscheidend. Der Austausch von Erfahrungen schafft Qualität, Praxisbezug und Innovation.
2. Ganzheitliche Perspektiven sind notwendig. Ökologische, soziale und ökonomische Aspekte der Digitalisierung müssen integriert gedacht werden.
3. Fehlende Sichtbarkeit ist ein zentrales Problem. Viele CDR-relevante Aktivitäten existieren bereits, werden aber nicht strategisch eingeordnet oder kommuniziert.

Hier setzt die CDR-Initiative an: Sie unterstützt Unternehmen mit klaren Methoden, Vernetzung und Orientierung – insbesondere durch ein Reifegradmodell, das den Status digitaler Verantwortung sichtbar macht, Entwicklungspfade aufzeigt und CDR mit bestehenden Digital- und Nachhaltigkeitsstrategien verknüpft, z. B. im ESG-Kontext.

3. Die Architektur von CDR

Corporate Digital Responsibility (CDR) basiert auf einer klaren und strukturierten Architektur, die Unternehmen eine Orientierung für eine verantwortungsvolle Digitalisierung bietet und deren „Meisterstück" das Reifegradmodell ist. Es besteht aus insgesamt sieben Elementen, die sich gegenseitig ergänzen und zusammen einen praktischen Rahmen für digitale Verantwortung bilden.

Neben dieser technischen Architektur der sieben Elemente gibt es eine ebenso zentrale, wenn auch oft implizite Dimension: die Governance des CDR. Sie umfasst die Steuerung, Koordination und Weiterentwicklung der Elemente als kohärentes Gesamtsystem – mit anderen Worten: die „Choreographie" des Reifeprozesses. In diesem Sinne bildet die Governance den organisatorischen Rahmen für eine verantwortungsvolle digitale Transformation.

1. Die neun ethischen „CDR-Prinzipien" bilden das normative Fundament. Sie definieren grundlegende Werte wie soziale Grundwerte, Fairness, Transparenz und Nachhaltigkeit und dienen als Leitlinien für digitale Unternehmensstrategien.
2. Der „CDR-Kodex" ist eine freiwillige Selbstverpflichtung für Unternehmen, die über gesetzliche Vorgaben hinausgeht. Er sorgt dafür, dass digitale Verantwortung nicht nur Theorie bleibt, sondern durch unternehmerisches Handeln sichtbar wird.
3. Um die Grundsätze in die Praxis umzusetzen, definiert der Kodex fünf zentrale „CDR-Handlungsfelder": (1) Umgang mit Daten, (2) Bildung, (3) Klima- und Ressourcenschutz, (4) Mitarbeitenden-Einbindung und (5) Inklusion. Jedes dieser Felder adressiert spezifische Herausforderungen der digitalen Transformation.
4. Die Handlungsfelder enthalten konkrete „CDR-Ziele", an denen sich Unternehmen bei ihren Maßnahmen orientieren können. Sie reichen von Datenschutz und algorithmischer Fairness über die Förderung der digitalen Teilhabe bis hin zu nachhaltigen IT-Strukturen.
5. Umgesetzt werden diese Ziele durch konkrete „CDR-Maßnahmen", die Unternehmen ergreifen können, um digitale Verantwortung zu übernehmen. Dazu gehören Schulungen zur digitalen Ethik, Strategien zur nachhaltigen IT-Nutzung und die transparente Gestaltung von Algorithmen.
6. Um ihre digitale Verantwortung zu bewerten und weiterzuentwickeln, können Unternehmen den „CDR-Quick-Check" nutzen, der eine erste Einschätzung ihres Stands ermöglicht.
7. Das „CDR-Reifegradmodell" schließlich hilft Unternehmen, ihre Fortschritte systematisch zu erfassen und weiterzuentwickeln.

Die sieben Elemente sind sinnvoll aufeinander abgestimmt und ergeben ein schlüssiges Gesamtkonzept, das über mehrere Jahre hinweg entwickelt wurde – im engen Dialog zwischen Politik und Wirtschaft. Die einzelnen Elemente werden nun im Detail erläutert.

Element 1: CDR-Prinzipien

Die neun Prinzipien bilden die ethische Grundlage für Entscheidungen und Handlungen. Sie geben die Richtung für eine verantwortungsvolle Digitalisierung vor. Die CDR-Initiative des BMUV hat die folgenden neun Prinzipien in einem Multi-Stakeholder-Prozess definiert (vgl. CDR-Initiative 2021).

1. Gesellschaftliche Grundwerte: Technik muss Demokratie, Freiheit, soziale Marktwirtschaft und Gleichbehandlung unterstützen.
2. Menschenzentrierung: Der Mensch steht im Mittelpunkt der digitalen Innovation.
3. Nutzen schaffen: Technologien müssen einen greifbaren Mehrwert bieten und ethisch vertretbar sein.
4. Schaden vermeiden: Digitale Systeme sollten sicher, kalkulierbar und frei von Risiken für die Verbraucher sein.
5. Autonomie: Verbraucher sollten selbstbestimmt über ihre Daten und digitalen Entscheidungen verfügen.
6. Fairness: Der Zugang zur digitalen Welt muss für alle offen und diskriminierungsfrei sein.
7. Transparenz: Die Unternehmen müssen klare Informationen über die Funktionsweise ihrer Technologien bereitstellen.
8. Verantwortlichkeit: Klar definierte menschliche Entscheidungsträger tragen die Verantwortung für digitale Prozesse.
9. Nachhaltigkeit: Die Digitalisierung soll ressourcenschonend sein und im Einklang mit den UN-Nachhaltigkeitszielen (Sustainable Development Goals) stehen.

Die Mitgliedsunternehmen verpflichten sich zur Einhaltung dieser Grundsätze, die als Leitlinien für den CDR-Kodex dienen.

Element 2: CDR-Kodex

Der CDR-Kodex ist eine freiwillige Selbstverpflichtung der Unternehmen, die über die gesetzlichen Anforderungen hinausgeht und ein klares Bekenntnis zur digitalen Verantwortung darstellt. Unternehmen, die diesen Kodex unterzeichnen, verpflichten sich freiwillig, die digitale Welt zum Wohle der Gesellschaft aktiv zu gestalten (vgl. ebd.). Der CDR-Kodex ist mehr als eine Ansammlung von guten Vorsätzen – er ist ein klarer Handlungsrahmen für eine digitale Zukunft, die auf

Verantwortung, Nachhaltigkeit und Fairness basiert. Das öffentliche Bekenntnis der unterzeichnenden Unternehmen schafft nicht nur Transparenz, sondern setzt auch ein starkes Signal für eine digitale Welt, die den Menschen in den Mittelpunkt stellt. Der CDR-Kodex bietet dafür eine wertvolle Grundlage – ein Versprechen für eine nachhaltige, faire und verantwortungsvolle digitale Zukunft. Die Mitgliedsunternehmen verpflichten sich zur Einhaltung des CDR-Kodex und veröffentlichen einen CDR-Bericht, in dem sie ihre Maßnahmen und Fortschritte im Bereich der digitalen Verantwortung darlegen (vgl. u.a. DKB 2023; ING-DiBa AG 2023; Otto GmbH & Co KG 2023 und Telefónica Deutschland Holding AG 2023). Konkretisiert wird der Kodex durch die CDR-Handlungsfelder.

Element 3: CDR-Handlungsfelder

Handlungsfelder sind thematische Bereiche, in denen spezifische Maßnahmen und Strategien umgesetzt werden, um übergeordnete Ziele zu erreichen. Sie dienen als operative Leitlinien, die Unternehmen oder Organisationen dabei helfen, ihre Grundsätze und Verpflichtungen in die Praxis umzusetzen. Die CDR-Initiative hat fünf Handlungsfelder definiert (vgl. CDR-Initiative 2021):

1. Umgang mit Daten: Datenschutz, ethische Datennutzung und Schutz vor algorithmischer Verzerrung.
2. Bildung: Stärkung der digitalen Kompetenzen, Sensibilisierung für Risiken und Förderung eines verantwortungsvollen Verhaltens.
3. Klima- und Ressourcenschutz: Entwicklung von umweltfreundlichen IT-Lösungen und Reduzierung des Energieverbrauchs digitaler Systeme.
4. Mitarbeitenden-Einbindung: Mitarbeiter auf die digitale Transformation vorbereiten und sie aktiv in Entscheidungsprozesse einbinden.
5. Inklusion: Sicherstellen, dass niemand von der Digitalisierung ausgeschlossen wird und alle von den technologischen Fortschritten profitieren.

Daten sind der Kernbestandteil der digitalen Verantwortung und bilden die Grundlage für alle anderen Bereiche der CDR. Sie sind das Rohmaterial, das den digitalen Wandel vorantreibt, und ihr verantwortungsvoller Umgang entscheidet über die Qualität und die Auswirkungen aller weiteren CDR-Bemühungen. Nur wenn Daten verantwortungsvoll verarbeitet werden, kann die digitale Verantwortung von Unternehmen nachhaltig und sozial wirksam umgesetzt werden. Jedes Handlungsfeld hat spezifische CDR-Ziele, an denen sich die Mitgliedsunternehmen orientieren.

Element 4: CDR-Ziele

In den jeweiligen Handlungsfeldern wurden gemeinsam mit den Mitgliedsunternehmen der CDR-Initiative in einem aufwändigen Abstimmungsprozess konkrete Ziele definiert. Die meisten Ziele wurden für den Umgang mit Daten übernommen, da dies das zentrale Handlungsfeld im Kodex ist (vgl. ebd.).

Ziele im Handlungsfeld 1: Umgang mit Daten

- Aktiv gegen Verzerrungen bei der Datenanalyse vorgehen
- Verantwortungsvolle, faire und transparente Profilerstellung
- Sicherstellung der Souveränität und Autonomie der Verbraucher
- Förderung einer verantwortungsvollen Gestaltung der Technologie im Umgang mit Daten
- Sicherstellung eines verantwortungsvollen Umgangs mit Daten im Unternehmen
- Übernahme von Verantwortung für den Umgang mit Daten über das eigene Unternehmen hinaus
- Systematische Verbesserung der Daten- und Cybersicherheit

Ziele im Handlungsfeld 2: Bildung

- Über Chancen und Risiken informieren und Verbraucher und Beschäftigte zu eigenverantwortlichem Handeln befähigen
- Informieren über ethische Fragen der Digitalisierung

Ziele im Handlungsfeld 3: Klima- und Ressourcenschutz

- Stärkung digitaler Lösungen zum Schutz unserer Umwelt
- Umwelt- und ressourcenschonende Gestaltung und Nutzung von Informations- und Kommunikationstechnologien

Ziele im Handlungsfeld 4: Mitarbeitenden-Einbindung

- Mitarbeiter in die Gestaltung und Entscheidungsfindung der digitalen Transformation einbeziehen
- Mitarbeiter auf den Wandel vorbereiten und bei der Transformation unterstützen

Ziele im Handlungsfeld 5: Inklusion

- Zugangsbarrieren zu Produkten und Dienstleistungen abbauen
- Förderung von Lösungen zur Unterstützung der Teilhabe

Die Ziele werden durch Maßnahmen umgesetzt, über die die Mitgliedsunternehmen berichten. Diese Maßnahmen gelten als Best Practice.

Element 5: CDR-Maßnahmen

CDR-Maßnahmen sind konkrete Unternehmensaktivitäten, die der Umsetzung der CDR-Ziele in den definierten Handlungsfeldern dienen. Während die CDR-Ziele den strategischen Rahmen vorgeben, sind die CDR-Maßnahmen die operativen Schritte, mit denen Unternehmen die digitale Verantwortung umsetzen. Im Folgenden wird jeweils ein Ziel pro Handlungsfeld mit einem Unternehmensbeispiel (vgl. u.a. DKB 2023; ING-DiBa AG 2023; Otto GmbH & Co KG 2023; Telefónica Deutschland Holding AG 2023 und Zalando SE 2023) dargestellt (vgl. Abbildung 1).

Die nachfolgenden Beispiele verdeutlichen, wie unterschiedlich die Ansätze der Mitgliedsunternehmen der CDR-Initiative sind, wenn es um die Förderung einer verantwortungsvollen und nachhaltigen Digitalisierung geht. Der CDR-Quick-Check und das CDR-Reifegradmodell sollen sicherstellen, dass sich die Unternehmen weiterentwickeln können. Damit werden ausdrücklich auch Unternehmen angesprochen, die in der CDR-Entwicklung noch nicht so weit fortgeschritten sind, aber bereit sind, sich in diesem Bereich stärker zu engagieren. In diesem Sinne versteht sich die CDR-Initiative des BMUV als Lehr- und Lernplattform.

CDR-Handlungsfeld	CDR-Ziel	CDR-Maßnahme eines Unternehmens
Umgang mit Daten	Daten- und Cybersicherheit konsequent weiterentwickeln	DKB: Einführung einer Informationssicherheitskampagne zu Daten- und Cybersicherheit
Bildung	Über ethische Fragen der Digitalisierung aufklären	ING: Anlaufstelle für Mitarbeitende für ethische Fragen rund um die Datenerhebung und -nutzung
Klima- und Ressourcenschutz	Informations- und Kommunikationstechnologien umwelt- und ressourcenschonend ausgestalten und nutzen	Otto Group: Umweltbewusste Softwareentwicklung durch Sustainable Programming
Mitarbeitenden-Einbindung	Mitarbeitende auf Veränderungen vorbereiten und im Wandel unterstützen	Telefónica: Lernprogramm „Digital Basics Learning Journey“, um Digitalkompetenzen zu vermitteln
Inklusion	Lösungen zur Unterstützung der Teilhabe fördern	Zalando: Anpassung der Online-Plattform zur Verbesserung des Zugangs für Menschen mit Behinderungen

ABBILDUNG 1: BEISPIELHAFTES ZUSAMMENSPIEL VON CDR-HANDLUNGSFELD, CDR-ZIEL UND CDR-MAßNAHME DER MITGLIEDSUNTERNEHMEN (QUELLE: EIGENE ABBILDUNG)

Element 6: Der CDR-Quick-Check

Der CDR-Quick-Check ist ein Selbstbewertungsinstrument, das Unternehmen dabei hilft, ihren aktuellen Status im Bereich der digitalen Verantwortung von Unternehmen (CDR) zu analysieren. Unternehmen können schnell und strukturiert feststellen, wie gut sie die digitale Verantwortung bereits umsetzen. Der Quick-Check ermöglicht eine erste Einschätzung der aktuellen Situation und identifiziert den Handlungsbedarf in den fünf CDR-Handlungsfeldern. Die Ergebnisse helfen, gezielte Maßnahmen abzuleiten, um die digitale Verantwortung systematisch weiterzuentwickeln. Der CDR-Quick-Check orientiert sich an den Handlungsfeldern und Zielen des CDR-Kodex. Der CDR-Quick-Check ist ein niedrigschwelliger Einstieg, um digitale Verantwortung systematisch zu erfassen und gezielt weiterzuentwickeln. Unternehmen erhalten einen klaren Überblick über Stärken und Entwicklungsfelder. Damit wird der ursprüngliche Gedanke, den CDR-Kodex lediglich als Berichtsinstrument zu sehen, zunehmend zur Grundlage für die aktive Übernahme digitaler Verantwortung. Der CDR-Quick-Check setzt den Kodex in die Praxis um und hilft Unternehmen, sich zu orientieren.

Element 7: Das CDR-Reifegradmodell

Darauf aufbauend ist das Reifegradmodell ein systematischer Bewertungsrahmen, mit dem Unternehmen ihren aktuellen Stand bei der Umsetzung von Corporate Digital Responsibility erfassen und weiterentwickeln können. Basierend auf dem CDR-Quick-Check hilft und motiviert es Unternehmen, digitale Verantwortung strukturiert und schrittweise in der Organisation zu verankern. Unternehmen können analysieren, wie weit sie mit der Umsetzung ihrer CDR-Strategie gekommen sind. Zudem können sie den aktuellen Stand des Transformationsprozesses dokumentieren, die interne Kommunikation erleichtern und die Akzeptanz in der Organisation erhöhen. Das Modell zeigt auch, welche Schritte notwendig sind, um die digitale Verantwortung gezielt auszubauen. Unternehmen können ihre „individuellen Prioritäten setzen", ohne jedes Thema auf den höchsten Reifegrad bringen zu müssen. Governance wird hier nicht als separate Funktion verstanden, sondern als integrativer Steuerungsmechanismus, der die strategische Einbettung und kontinuierliche Verbesserung von CDR-Prozessen ermöglicht. Sie verbindet normative Grundsätze mit der organisatorischen Umsetzung – und übersetzt Werte in die operative Realität. Unser Reifegradmodell basiert auf dem Capability Maturity Model Integration (CMMI) und besteht aus fünf Reifegradstufen. Das CMMI ist ein weltweit anerkanntes Modell, das Organisationen hilft, ihre Prozesse zu verbessern und effizienter zu gestalten (vgl. CMMI 2025) (vgl. Abbildung 2):

Reifegrad	Beschreibung
Initial	CDR-Themen werden vereinzelt aufgegriffen, aber noch nicht systematisch bearbeitet.
Projektiert	Erste Strukturen entstehen, die Organisation experimentiert mit CDR-Ansätzen. Temporäre Widersprüche mit anderen Unternehmenszielen sind noch akzeptabel.
Definiert	Die CDR-Ziele sind klar formuliert, Aufgaben sind zugeordnet und unternehmensweit kommuniziert.
Gesteuert	Der Fortschritt wird regelmäßig überprüft, es gibt Kennzahlen zur Erfolgsmessung.
Optimiert	Die CDR-Maßnahmen werden kontinuierlich verbessert und haben eine hohe Priorität in der Unternehmensstrategie.

ABBILDUNG 2: DAS REIFEGRADMODELL
(QUELLE: EIGENE ABBILDUNG)

4. Die Philosophie von Reifegradmodellen und ihre Konkretisierung

Der philosophische Hintergrund von Reifegradmodellen findet sich bereits in Aristoteles' Vorstellung von der Entwicklung des menschlichen Potenzials und dem Capability-Ansatz von Amartya Sen und Martha Nussbaum. Aristoteles versteht Reife als einen Prozess, in dem sich ein Wesen seinem inneren Ziel („telos") nähert – durch Übung, Tugend und rationales Handeln erreicht der Mensch ein gelingendes Leben („eudaimonia") (vgl. Aristoteles 1999). Sen und Nussbaum greifen diesen Gedanken auf und entwickeln ihn weiter zu der Frage: Was braucht der Mensch, um ein wirklich freies und selbstbestimmtes Leben führen zu können? Ihr Fokus liegt nicht auf bloßen Ergebnissen, sondern auf den realen Fähigkeiten, d.h. darauf, welche Chancen und Möglichkeiten ein Mensch tatsächlich hat, sein Potenzial zu verwirklichen (vgl. Sen 1999, Nussbaum 2011). Übertragen auf Organisationen und technische Systeme bedeutet dies, dass sich Reife nicht nur in Leistung oder Struktur widerspiegelt, sondern auch in der wachsenden Fähigkeit, sinnvolle Ziele zu verfolgen, Verantwortung zu übernehmen und Menschen zu befähigen. In diesem Sinne sind Reifegradmodelle mehr als nur Instrumente zur Effizienzsteigerung – sie orientieren sich an der Qualität der Entwicklung im Sinne von Potenzialentfaltung und ethischer Verantwortung.

Diese Vorstellung lässt sich direkt auf Organisationen und technische Systeme übertragen: Auch sie können reifen, wenn sie lernen, reflektieren und Verantwortung übernehmen. Technische Systeme, etwa im Zuge der Digitalisierung, entwickeln sich nicht bloß funktional weiter,

sondern durchlaufen Phasen, in denen sich ihre Reife daran bemisst, wie gut sie menschliche Bedürfnisse unterstützen (vgl. Brink 2025). Unser Reifegradmodell ist vor diesem Hintergrund zu verstehen: Es zielt nicht nur auf technologische Exzellenz oder Prozessqualität, sondern bewertet, inwieweit digitale Innovation im Einklang mit menschlicher Entwicklung, gesellschaftlichen Werten und ökologischer Verantwortung gestaltet wird. Es verbindet also technische Systemlogik mit ethischer Orientierung.

Reifegradmodelle haben ihre Ursprünge im Qualitäts- und Prozessmanagement. Das „Quality Management Maturity Grid" von Crosby (1979) gilt als eines der ersten Modelle zur stufenweisen Bewertung organisationaler Entwicklung. Das später entwickelte Capability Maturity Model (CMM) und seine Weiterentwicklung CMMI (vgl. CMMI 2025) fanden breite Anwendung, insbesondere im IT-Bereich. Reifegradmodelle strukturieren komplexe Veränderungsprozesse in nachvollziehbare Stufen und übernehmen dabei drei zentrale Funktionen: Sie erfassen systematisch den Ist-Zustand (deskriptiv), geben konkrete Handlungsempfehlungen (präskriptiv) und ermöglichen organisationsübergreifende Vergleichbarkeit (kompensatorisch) (vgl. Pöppelbuß/Röglinger 2011). Ein Beispiel ist der Industrie 4.0 Maturity Index, der die digitale Transformation industrieller Prozesse systematisiert (vgl. acatech 2020).

Ein zentraler Vorteil von Reifegradmodellen ist ihre Anpassungsfähigkeit. Sie lassen sich auf verschiedene Branchen, Unternehmensgrößen und strategische Zielsetzungen zuschneiden. Gerade im vielfältigen Feld der Corporate Digital Responsibility – mit Handlungsfeldern wie Datenschutz, Klimaschutz oder Inklusion – bieten sie einen strukturierenden Rahmen, der sich gut in bestehende Managementsysteme integrieren lässt. Erfolgreiche Reifegradmodelle verknüpfen unterschiedliche Themenfelder systematisch und fördern so ganzheitliches Lernen. Sie ermöglichen kontinuierliche Selbstreflexion und nachhaltige Weiterentwicklung entlang definierter Kriterien. Im Rahmen von Netzwerkinitiativen wie der CDR-Initiative dienen sie zusätzlich als gemeinsame Referenz und Kommunikationsbasis. Insbesondere in Entwicklungspartnerschaften bieten sie eine gemeinsame Sprache zur praxisnahen Weiterentwicklung von CDR-Maßnahmen – auch im Hinblick auf konkrete Herausforderungen und Lösungsstrategien (vgl. Lasrado 2018; Bititci et al. 2015; Dijkman et al. 2016).

5. Ein CDR-Reifegradmodell und seine Anwendungen

Das in der CDR-Initiative entwickelte Reifegradmodell hat sich in verschiedenen Anwendungskontexten als stabil und tragfähig erwiesen. In seinem Zentrum stehen die Prinzipien, Handlungsfelder und Ziele der CDR-Initiative, die trotz ihrer Flexibilität eine konsistente Orientierung bieten. Die konzeptionelle Stärke des Modells liegt darin, dass CDR nicht als isoliertes Thema, sondern als integrativer Bestandteil des unternehmerischen Handelns verstanden wird. CDR ist keine separate Funktion, sondern bildet die Grundlage für verantwortliches, nachhaltiges Organisationshandeln in seiner Gesamtheit. Ein zentrales Entwicklungsprinzip war: Keine Entwicklung ohne Nutzenperspektive. Für jede Reifegradstufe wurde daher ein konkreter Mehrwert formuliert, der sich in der Praxis direkt realisieren lässt, z. B. durch erhöhte Prozesssicherheit, bessere Entscheidungsfähigkeit, regulatorischen Schutz oder verbesserte Wirkung auf Stakeholder. Diese Nutzenorientierung hat sich als besonders geeignet für Organisationen erwiesen, die CDR nicht aus Compliance-Gründen, sondern aus strategischer Überzeugung betreiben. Ausgehend von den bisherigen Erfahrungen mit Entwicklungspartnern lassen sich drei typische Phasen unterscheiden:

Orientierung schaffen durch den CDR-Quick-Check

Der CDR-Quick-Check ist der niedrigschwellige Einstieg. In ca. 10-20 Minuten erhalten die Organisationen eine erste Einschätzung ihrer digitalen Verantwortungspraxis entlang zentraler Handlungsfelder. Das Tool ist bewusst schlank gehalten und liefert eine erste Einschätzung von Stärken, Schwächen und Entwicklungspotenzialen. Es eignet sich besonders zur internen Selbstreflexion und Sensibilisierung sowie zur Vorbereitung eines Workshops, z. B. zur strategischen Positionierung oder Priorisierung von Maßnahmen.

Etablierung einer unbürokratischen Lean Governance

Wird das CDR weiterverfolgt, empfiehlt sich eine schlanke Governance-Struktur. Vier pragmatische Schritte haben sich in mehreren Entwicklungspartnerschaften bewährt:

1. Verankerung: Explizite Einbeziehung der digitalen Verantwortung in die Vision und Strategie, z. B. durch digitale Leitlinien.
2. Fokussierung: Festlegung von Prioritäten für die wichtigsten Handlungsfelder.

3. Bestandsaufnahme: Detaillierte Analyse des Status quo und Entwicklung eines Perspektivplans.
4. Unterstützung des Wandels: Planung von Kommunikations- und Beteiligungsformaten zur Unterstützung des Kulturwandels.

Diese Schritte können mit überschaubarem Aufwand realisiert werden. Die Etablierung einer schlanken Governance-Struktur ist ein wichtiger Entwicklungsschritt in diesem Prozess: Sie schafft organisatorisches Commitment und die notwendige Abstimmung zwischen den Elementen der CDR-Architektur. Governance bedeutet hier nicht Bürokratie, sondern sinnvolle Verankerung, Reflexion und Zuweisung von Verantwortung.

Integration: Vom Projekt zur Praxis

Nach der anfänglichen Umsetzung und den ersten Maßnahmen etabliert sich das CDR zunehmend als Teil regulärer Strukturen. Organisationen wie die Deutsche Telekom zeigen, wie das Thema nach einer Projektphase in bestehende Prozesse, Rollen und Routinen integriert werden kann. Dabei geht es nicht nur um Regeln und Vorschriften, sondern auch um Kulturarbeit: CDR wird Teil der gemeinsamen Ausrichtung, z. B. in der Produktentwicklung, der Personalentwicklung oder dem Stakeholder-Dialog.

6. Fazit und Ausblick

Ein wesentlicher Erfolgsfaktor des CDR-Reifegradmodells liegt in seiner konsequenten Orientierung an der Unternehmenspraxis. Die CDR-Initiative hat in den letzten Jahren ein breites Portfolio an Fallbeispielen – von kleinen und mittleren Unternehmen bis hin zu Großkonzernen – aufgebaut, das nicht nur Orientierung bietet, sondern auch den Austausch zwischen Organisationen ermöglicht. Diese praktischen Referenzen fördern das Peer-Learning, regen zum Nachdenken an und schaffen Räume für einen konstruktiven Dialog über Unternehmensverantwortung im digitalen Zeitalter.

Das CDR-Reifegradmodell verbindet Anspruch mit Alltagstauglichkeit: Es macht digitale Verantwortung greifbar, ohne Organisationen zu überfordern. Die drei Entwicklungsstufen bieten einen niederschwelligen, aber strukturierten Einstieg in die Umsetzung von CDR. Sie fördern

systematisches Lernen, schaffen einen gemeinsamen Bezugsrahmen und stärken die organisationsübergreifende Vernetzung – besonders relevant in Zeiten des doppelten Wandels, in denen ökologische und digitale Anforderungen immer stärker miteinander verwoben sind. Gleichzeitig ist es wichtig, selbstkritisch zu reflektieren: Reifegradmodelle sind keine Patentlösungen. Sie können helfen, Orientierung zu geben und Fortschritte sichtbar zu machen, aber sie laufen Gefahr, komplexe Entwicklungen zu vereinfachen oder in Checklisten zu erstarren. Der eigentliche Mehrwert liegt nicht im Modell selbst, sondern in seiner lebendigen Anwendung – im Dialog, in der Anpassung an den jeweiligen Kontext und im ständigen Hinterfragen der eigenen Positionen.

Die nächsten Jahre werden zeigen, ob und wie sich das CDR-Reifegradmodell als dauerhaftes Managementinstrument etabliert. Eines ist klar: Die Anforderungen an die unternehmerische Verantwortung in der digitalen Transformation steigen weiter – aber auch die Instrumente entwickeln sich weiter. Das Modell der CDR-Initiative bietet einen praxisnahen, anpassungsfähigen Rahmen, der offen ist für Weiterentwicklungen und neue Anwendungsszenarien. In Zukunft wird es entscheidend sein, die Governance der CDR-Initiative sowie ihre inhaltliche Architektur weiterzuentwickeln. Nur wenn Steuerungsprozesse, Lernzyklen und normative Vorgaben gleichermaßen berücksichtigt werden, kann Corporate Digital Responsibility ihre volle Wirkung entfalten.

Literaturverzeichnis

acatech (2020): Industrie 4.0 Maturity Index – Die digitale Transformation von Unternehmen gestalten, München: Herbert Utz Verlag.

Accenture (2021): The European Double Up. A Twin Strategy That Will Strengthen Competitiveness, URL: https://www.accenture.com/content/dam/accenture/final/a-com-migration/r3-3/pdf/pdf-144/accenture-the-european-double-up.pdf (aufgerufen am: 29/06/2025).

Aristotle (1999): Nicomachean Ethics. Übers. v. Terence Irwin, Indianapolis: Hackett Publishing.

Bititci, U. S. / Bourne, M. / Cross, J. A. / Nudurupati, S. S. / Sang, K. (2015): Towards a Theoretical Foundation for Performance Measurement and Management, in: International Journal of Management Reviews, Jg. 17 / Nr. 4, 534–536, DOI: https://doi.org/10.1111/ijmr.12086.

Brink, A. (2022): Digitale Ethik, in: Aßländer, M. S. (Hrsg.): Handbuch Wirtschaftsethik, 2. Aufl, Stuttgart: Metzler, 615–624.

– (2025): Menschenzentrierte Digitalisierung, Baden-Baden: Nomos.

CDR-Initiative (2021): Code of Corporate Digital Responsibility. Voluntary compliance including reporting, URL: https://cdr-initiative.de/uploads/files/2024-01_Code_CDRInitiative_EN.pdf (aufgerufen am: 29/06/2025).

– (2025a): Corporate Digital Responsibility Berichte, URL: https://cdr-initiative.de/cdr-berichte (accessed: 29/06/2025).

– (2025b): Corporate Digital Responsibility Initiative, URL: https://cdr-initiative.de/initiative (accessed: 29/06/2025).

Christ, A.-S. (Hrsg.) (2025): Digital Responsibility: From Principles to Practice, Berlin: Springer [im Erscheinen].

CMMI – Capability Maturity Model Integration (2025): Capability Maturity Model Integration Institut, URL: https://cmmiinstitute.com/ (aufgerufen am: 29/06/2025).

Crosby, P. B. (1979): Quality is Free: The Art of Making Quality Certain, New York: McGraw-Hill.

Dijkman, R. / Lammers, S. V. / de Jong, A. (2016): Properties that Influence Business Process Management Maturity and its Effect on Organizational Performance, in: Information Systems Frontiers, Jg. 18 / Nr. 4, 717–734, DOI: https://doi.org/10.1007/s10796-015-9554-5.

DKB AG (2023): CDR-Kodex Maßnahmenbericht (DKB AG), URL: https://cdr-initiative.de/uploads/files/Berichte/DKB-Report-2023.pdf (aufgerufen am: 29/06/2025).

ING-DiBa AG (2023): CDR-Kodex Maßnahmenbericht (ING-DiBa AG), URL: https://cdr-initiative.de/uploads/files/Berichte/ING-Report-2023.pdf (aufgerufen am: 29/06/2025).

Lasrado, L. A. (2018): Set Theoretic Approach to Maturity Models. Dissertation, Copenhagen Business School, PhD Series Nr. 15.2018.

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: https://doi.org/10.1016/j.jbusres.2019.10.006.

Merbecks, U. (2023): Corporate Digital Responsibility (CDR) in Germany: Background and First Empirical Evidence from DAX 30 Companies in 2020, in: Journal of Business Economics, Jg. 94 / Nr. 7–8, 1025–1049, DOI: https://doi.org/10.1007/s11573-023-01148-6.

Nussbaum, M. C. (2011): Creating Capabilities: The Human Development Approach, Cambridge: Harvard University Press.

Otto GmbH & Co KG (2023): CDR-Kodex Maßnahmenbericht (Otto [GmbH & Co KG]), URL: https://cdr-initiative.de/uploads/files/Berichte/Otto-Group-Report-2023.pdf (aufgerufen am: 29/06/2025).

Pöppelbuß, J. / Röglinger, M. (2011): What Makes a Useful Maturity Model? A Framework for General Design Principles for Maturity Models and its Demonstration in Business Process Management, in: Tuunainen, V. (.): Proceedings of the 19th European Conference on Information Systems (ECIS 2011), Helsinki, 9.–11. Juni 2011.

Sen, A. (1999): Development as Freedom, New York: Alfred A. Knopf.

Telefónica Deutschland Holding AG (2023): CDR-Kodex Maßnahmenbericht (Telefónica Deutschland Holding AG), URL: https://cdr-initiative.de/uploads/files/Berichte/Telefonica-Report-2023.pdf (aufgerufen am: 29/06/2025).

Zalando SE (2023): CDR-Kodex Maßnahmenbericht (Zalando SE), URL: https://cdr-initiative.de/uploads/files/Berichte/Zalando-Report-2022.pdf (aufgerufen am: 29/06/2025).

2.

Verbraucher-perspektive

Chancen und Risiken beim Einsatz von KI aus Sicht der Bevölkerung[1]

Ergebnisse einer repräsentativen Online-Befragung

Sara Elisa Kettner, Christian Thorun und Annette Cerulli-Harms

1. Einleitung

Die Chancen und Risiken künstlicher Intelligenz (KI) für die Gesellschaft werden breit und ambivalent diskutiert und liegen somit zentral im Wirkungsfeld der Corporate Digital Responsibility (CDR). Der CDR-Kodex der CDR-Initiative enthält handlungsleitende Prinzipien[2] und Zielsetzungen, zu denen sich die Mitglieder der CDR-Initiative bekennen. Nachdem auf der CDR-Konferenz im Jahr 2023 das Prinzip „Menschenzentrierung“[3] im Fokus stand (vgl. Brink 2024; Kettner/Thorun 2024), zielte die 5. CDR-Konferenz der CDR-Initiative am 14.11.2024 in Berlin auf das Prinzip „Fairness“[4]. Ziel war anhand konkreter Beispiele aufzuzeigen, wie das Fairness-Prinzip in der Entwicklung und beim Einsatz digitaler Lösungen, insbesondere bei KI-Anwendungen, berücksichtigt werden kann und sollte. Für einen erfolgreichen Einsatz von KI ist es essenziell, das Vertrauen der Bürger*innen zu fördern und deren Erwartungen an KI-Anwendungen zu verstehen.

---

1 ChatGPT-5 von OpenAI kam zum Einsatz für die Erstellung von Alternativtexten.

2 Der CDR-Kodex enthält insgesamt neun Prinzipien, welche abstrahiert den Zielen der fünf Handlungsfelder – Umgang mit Daten, Bildung, Klima- und Ressourcenschutz, Mitarbeitendeneinbindung und Inklusion – zugrunde liegen.

3 Prinzip 2: Menschenzentrierung: Wir stellen den Menschen bei der Entwicklung und dem Einsatz technischer Systeme in den Mittelpunkt (vgl. CDR-Initiative 2025b).

4 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 (vgl. CDR-Initiative 2025b).

Vor diesem Hintergrund hat die Geschäftsstelle der CDR-Initiative im Oktober 2024 eine online-repräsentative Befragung mit N=1.022 Befragten zu Erfahrungen mit KI-Anwendungen, den antizipierten Chancen und Risiken beim Einsatz von KI sowie Erwartungen an einen verantwortungsvollen KI-Einsatz durch Unternehmen durchgeführt. Die Ergebnisse werden im Folgenden dargestellt.

2. Ausgangslage

Im Alltag der Bürger*innen findet sich künstliche Intelligenz (KI) zunehmend sowohl in direkter als auch in indirekter Form in diversen Anwendungen. In direkter Form nutzten im Jahr 2023 bspw. 18% der Deutschen ein Tool mit generativer KI und immerhin 79% hatten bereits von solchen KI-Assistenten gehört (vgl. BCG 2024). Zudem finden KI-Systeme auch indirekt Einzug in den Konsumalltag, bspw. in Form von personalisierten Empfehlungsalgorithmen, zur Steuerung von Smart-Home-Anwendungen oder in Chatbots im Kundenservice.

Im Hinblick auf den Einsatz von KI gibt es aus Bürger*innensicht sowohl Chancen als auch Risiken. So erleichtern KI-Systeme den Alltag, können bei Konsumentscheidungen helfen und ermöglichen personalisierte Dienstleistungen. In einer Befragung aus dem Jahr 2023 gab die Mehrheit der Befragten (52%) an, dass sie in generativer KI das Potenzial sehen, in vielen Bereichen zu unterstützen und helfen zu können (vgl. Kero et al. 2023). Auf der anderen Seite können Abhängigkeiten entstehen, Datenschutz und Privatsphäre können beeinträchtigt werden und Biases in den zugrundeliegenden Datensätzen können zu Diskriminierung führen. Aus diesem Grund sind ganzheitliche Ansätze, um KI verantwortungsvoll und im Sinne der Bürger*innen zu gestalten, notwendig. Hierzu zählen übergeordnet Themen wie Transparenz und Erklärbarkeit, aber auch konkrete Anforderungen an Unternehmen, die KI einsetzen, sowie an die Politik, die die Risiken reguliert.

Bisher ist jedoch nicht ausreichend untersucht, welche konkreten Chancen und Risiken Bürger*innen in der Nutzung von KI-Systemen sehen und inwieweit bestimmte Anforderungen mit Blick auf einen verantwortungsvollen Einsatz von KI erfüllt sein sollten.

3. Befragungsmethodik

3.1 Zielsetzung und Fragenkatalog

Diese Forschungslücke schließt dieser Beitrag und berichtet die Ergebnisse einer Online-Umfrage zu den Erfahrungen mit und Erwartungen an KI. Insgesamt werden dabei drei Themenfelder behandelt:

1. Erfahrungen mit KI-Anwendungen,
2. Chancen und Risiken beim Einsatz von KI, und
3. Erwartungen der Bürger*innen an einen verantwortungsvollen Einsatz von KI.

Diese Themenkomplexe werden sowohl allgemein als auch mit Blick auf Unternehmen sowie die Politik untersucht.[5] Zudem wurden sozio-demografische Merkmale erhoben, um gruppenspezifische Analysen zu ermöglichen.

3.2 Stichprobe

Die verwendeten Daten beruhen auf einer Online-Umfrage, die von dem Marktforschungsdienstleister YouGov Deutschland GmbH durchgeführt wurde. An der Umfrage nahmen 1.022 Personen zwischen dem 14. und 17.10.2024 teil.[6] Die Ergebnisse wurden gewichtet und sind repräsentativ für die Bevölkerung in Deutschland ab 18 Jahren.[7] Weitere Informationen zur Stichprobe finden sich im Anhang zur Originalstudie.[8]

---

5 Der vollständige Befragungsbogen kann auf der Webseite der CDR-Initiative heruntergeladen werden (vgl. CDR-Inititative 2025a).

6 Hierfür wurde das YouGov-Panel genutzt. Die Panelist*innen wurden per E-Mail mit einem Link zur Umfrage eingeladen. Dabei erfolgte die Stichprobenziehung per Zufall aus der Grundgesamtheit des YouGov-Panels Deutschland gemäß den Vorgaben der Studie (hier: „deutsche Bevölkerung ab 18 Jahren").

7 Die resultierende Stichprobe wurde bei der Auswertung zusätzlich zur Quotierung gewichtet, um eine bevölkerungsrepräsentative Zielverteilung gewährleisten zu können. Bei der Quoten-Definition und Gewichtung wurden die Daten des Mikrozensus zugrunde gelegt (ausgehend von den Merkmalen Alter, Geschlecht und Region). Wie in der vorhergehenden Erläuterung genannt, basieren die Ergebnisse auf einer Zufallsstichprobe und erfüllen somit die Voraussetzungen für die Berechnung statistischer Kenngrößen und Tests.

8 Anhang 1: Stichprobenübersicht, Anhang 2: Ausgewählte Auswertungen nach sozio-demografischen Variablen sowie Anhang 3: Befragungsdesign. Die Anhänge können auf der Webseite der CDR-Initiative heruntergeladen werden (vgl. CDR-Inititative 2025a).

Die Antworten wurden mit Hilfe einer statistischen Auswertung durch das ConPolicy – Institut für Verbraucherpolitik analysiert und geben Auskunft über die jeweiligen Fragestellungen sowie hinsichtlich unterschiedlicher sozio-demografischer Aspekte. Hierzu zählen Alter, Geschlecht und digitale Affinität.[9] Die Anteile in Abbildungen, Tabellen und im Text sind für die bessere Lesbarkeit der Ergebnisse auf ganze Zahlen gerundet.

4. Ergebnisse

4.1 Erfahrungen mit KI-Anwendungen

Die Befragten sollten zuerst angeben, ob sie bereits „Erfahrungen mit der Nutzung von KI-Assistenten" wie bspw. OpenAI Chat-GPT, Google Gemini oder Microsoft Copilot gemacht haben. Dabei wurden fünf unterschiedliche Nutzungsfälle unterschieden, die in Abbildung 1 dargestellt sind.

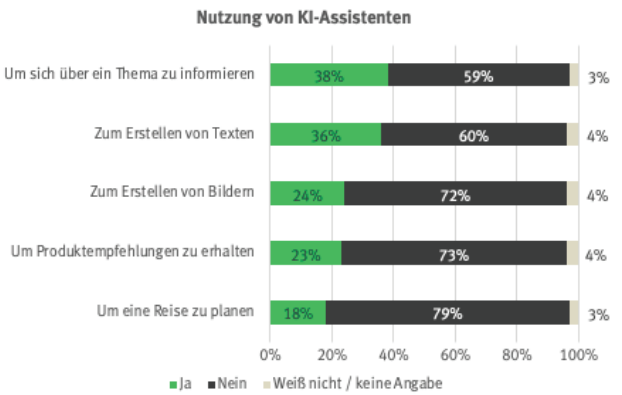


ABBILDUNG 1: NUTZUNG VON KI-ASSISTENTEN (QUELLE: EIGENE DARSTELLUNG)

[9] Bei den separat berechneten Analysen werden folgende Gruppen unterschieden: Alter (18 bis 24 Jahre, 25 bis 34 Jahre, 35 bis 44 Jahre, 45 bis 54 Jahre vs. 55 Jahre und älter), Geschlecht (männlich vs. weiblich), und digitale Affinität (niedrig, mittel vs. hoch). Es werden lediglich statistisch signifikante Ergebnisse berichtet. Weitere Angaben zur Zusammensetzung der Stichprobe finden sich im separaten Anhang 1: Stichprobenübersicht (vgl. CDR-Initiative 2025a).

Mit 38% und damit am häufigsten gaben die Befragten an, dass sie bereits einen KI-Assistenten genutzt haben, um sich über ein Thema zu informieren (59% keine Nutzung, 3% keine Angabe). An zweiter Stelle folgt das Erstellen von Texten mit Hilfe eines KI-Assistenten mit 36% (60% keine Nutzung, 4% keine Angabe). Zum Erstellen von Bildern haben bereits 24% der Befragten einen KI-Assistenten genutzt (72% keine Nutzung, 4% keine Angabe), gefolgt von 23% die angaben, dass sie einen KI-Assistenten bereits genutzt haben, um Produktempfehlungen zu erhalten (73% keine Nutzung, 4% keine Angabe). Zudem gaben 18% der der Befragten an, dass sie bereits mit Hilfe eines KI-Assistenten eine Reise geplant haben (79% keine Nutzung, 3% keine Angabe).

Darüber hinaus steht das bisherige Nutzungsverhalten der KI-Anwendungen teilweise im Zusammenhang mit sozio-demografischen Attributen. So lassen sich im Hinblick auf das Alter sowie die digitale Affinität Unterschiede feststellen (vgl. Tabellen im Anhang der Originalstudie).[10] Je älter die Befragten, desto seltener haben sie die unterschiedlichen KI-Anwendungen bisher genutzt. Je digital affiner die Befragten, desto häufiger nutzten sie die KI-Anwendungen bereits.

Anschließend wurden die Befragten nach ihren „Erfahrungen mit KI bei der Nutzung von Online-Diensten" sowie in sozialen Medien befragt.[11] „Fairness" im CDR-Kodex bedeutet, dass bei der Entwicklung und dem Einsatz technischer Systeme die Teilhabe und der Zugang zu einer digitalisierten Welt gestärkt werden soll. Dafür wurden in diesem Frageblock vier unterschiedliche Aspekte beleuchtet, die teilweise unvereinbar mit dem Fairness-Prinzip wären bzw. die Berücksichtigung des Prinzips unterstreichen. Die Antworten befinden sich in Abblidung 2:

---

10 Zur Untersuchung etwaiger Gruppenunterschiede wurden χ^2-Tests gerechnet bzw. im Falle von Zusammenhängen zwischen Attributen Spearman-Korrelationen. Es werden hier und im Folgenden nur statistisch signifikante Ergebnisse mit $p<0{,}05$ berichtet. Zudem konnte ein einzelner Unterschied zwischen Männern und Frauen festgestellt werden, nämlich bei der Erstellung von Bildern, die häufiger von Männern als von Frauen vorgenommen werden. Die weiteren Nutzungsarten stehen jedoch nicht in einem systematischen Zusammenhang mit dem Geschlecht.

11 Hierbei ist anzumerken, dass die vier genannten Beispiele, nämlich Falschinformationen, Chatbots, Hassrede und KI-Influencer*innen oftmals, aber nicht ausschließlich von KI-Systemen ausgeführt bzw. produziert werden.

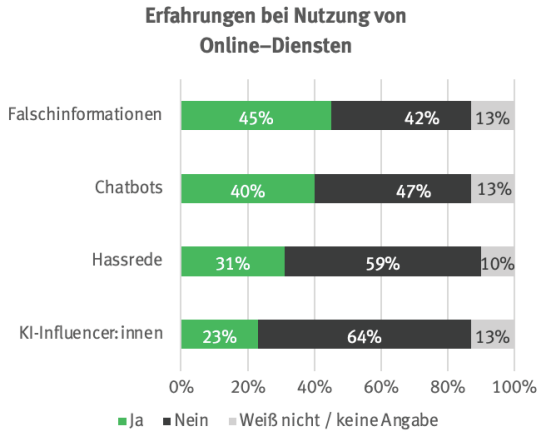


ABBILDUNG 2: ERFAHRUNGEN BEI NUTZUNG VON ONLINE-DIENSTEN (QUELLE: EIGENE DARSTELLUNG)

45% der Befragten gaben an, dass sie bei der Nutzung von Online-Diensten bereits Falschinformationen begegnet sind (42% nicht, 13% keine Angabe). 40% waren bereits Chatbots aufgefallen (47% nicht, 13% keine Angabe). Knapp ein Drittel (31%) berichtete, dass ihnen bereits Hassrede begegnet ist (59% bisher nicht, 10% keine Angabe). Den geringsten Anteil erhielten die KI-Influencer*innen, denen bisher jedoch immerhin 23% der Befragten aufgefallen waren (64% nicht aufgefallen, 13% keine Angabe). Auch in diesem Fall konnten in der Auswertung Unterschiede zwischen den sozio-demografischen Gruppen festgestellt werden, die in den Tabellen im separaten Anhang zu finden sind. Zum einen konnte ein negativer Zusammenhang zwischen Alter und berichteter Erfahrung festgestellt werde, d.h. je älter die Befragten, desto seltener gaben sie an, dass ihnen Falschinformationen, Chatbots, Hassrede und KI-Influencer*innen bereits begegnet waren. Die digitale Affinität steht hingegen in einem positiven Zusammenhang mit der berichteten Erfahrung, d.h. je digital affiner, desto häufiger waren ihnen auch die vier Aspekte bei der Nutzung von Online-Diensten aufgefallen. Die Unterschiede zwischen den Geschlechtern sind zwar nicht systematisch, jedoch gaben Männer häufiger als Frauen an, dass ihnen in Online-Diensten bereits Chatbots sowie KI-Influencer*innen aufgefallen waren.

4.2 Chancen und Risiken beim Einsatz von KI

Mit Blick auf den Einsatz von KI bestehen sowohl Chancen als auch Risiken für Bürger*innen. Während die Chancen die „Fairness" fördern können, wird das Prinzip durch die Risiken geschwächt. Diese Chancen und Risiken sind sehr vielfältig, sodass es von Interesse ist, sie einzeln bewerten zu lassen und so ihre Relevanz auch für das Prinzip „Fairness" abzuschätzen.

Chancen

Insgesamt wurden den Befragten fünf (potenzielle) Chancen, die sich aus der KI-Nutzung ergeben, vorgestellt. Dies waren a) personalisierte Empfehlungen, bspw. für Reisen, Filme oder Produkte, die hierdurch den Komfort der Bürger*innen erhöhen, b) eine Erleichterung durch Automatisierung, z. B. im Rahmen einer automatischen Heizungs- oder Beleuchtungssteuerung im Smart-Home-Bereich, c) die bequeme Erstellung von Texten, Bildern oder Videos mit generativen KI-Systemen, d) bessere Entscheidungen durch datenbasierte Analysen, wie bspw. bessere ärztliche Diagnosen, sowie e) der Anstoß technischer Innovationen durch KI, wodurch die Wettbewerbsfähigkeit des Standorts Deutschland gestärkt wird.

Insgesamt lässt sich feststellen, dass eine Mehrheit zwischen 61% und 77% den potenziellen Chancen (eher oder sogar voll und ganz) zustimmen. Zwischen den einzelnen Chancen gibt es dabei gewisse Unterschiede. So stimmten 70% der Befragten zu, dass KI-Systeme mit Hilfe von personalisierten Empfehlungen den Komfort erhöhen (21% stimmten nicht zu, 9% keine Angabe). 77% stimmten zu, dass KI-gesteuerte Automatisierungen dabei helfen können, den Alltag zu erleichtern (17% stimmten nicht zu, 6% keine Angabe). Zudem stimmten 73% zu, dass aus ihrer Sicht mit generativen KI-Systemen bequem Inhalte erstellt werden können (18% stimmten nicht zu, 9% keine Angabe). 68% stimmten zudem zu, dass durch datenbasierte Analysen bessere Entscheidungen getroffen bzw. Diagnosen gestellt werden können (20% stimmten nicht zu, 12% keine Angabe). Zuletzt stimmten 61% zu, dass durch KI-Systeme technische Innovationen angestoßen werden, die die Wettbewerbsfähigkeit stärken können (24% stimmten nicht zu, 15% keine Angabe). Abbildung 3 zeigt, inwiefern die Befragten den fünf potenziellen Chancen zustimmen.

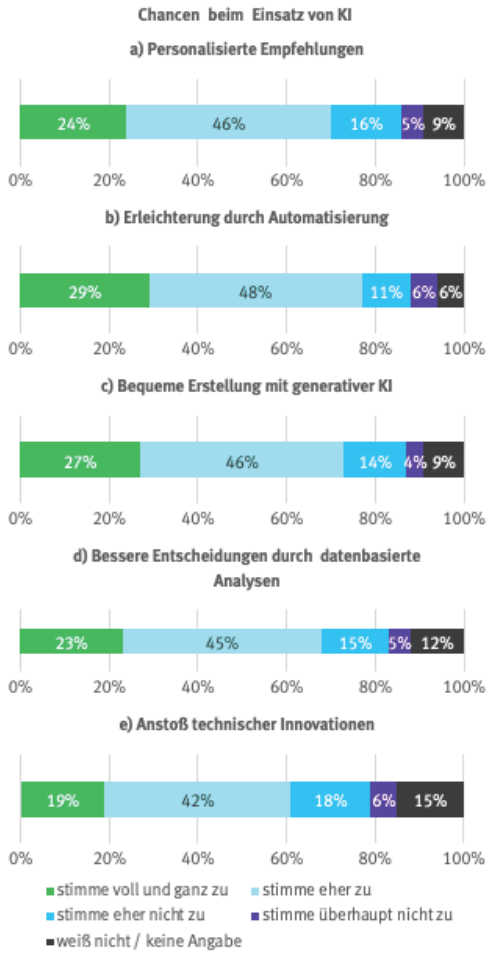


ABBILDUNG 3: CHANCEN BEIM EINSATZ VON KI
(QUELLE: EIGENE DARSTELLUNG)

Mit Blick auf sozio-demografische Variablen lassen sich auch hier Unterschiede feststellen, die im separaten Angang dargelegt sind. Besonders sticht hierbei der Zusammenhang mit der digitalen Affinität der Befragten hervor. So lassen sich für alle fünf potenziellen Chancen positive, signifikante Zusammenhänge feststellen, die zeigen, dass je digital affiner die Befragten sind, desto

mehr erwarten sie auch die positiven Effekte bzw. Chancen durch den Einsatz der KI. Im Hinblick auf das Alter und das Geschlecht der Befragten lassen sich zwar keine systematischen Unterscheide feststellen, jedoch für einige ausgewählte Chancen. So besteht ein schwacher, signifikant negativer Zusammenhang zwischen dem Alter und den antizipierten Chancen von personalisierten Empfehlungen sowie der Erleichterung durch Automatisierung. Dies weist darauf hin, dass je älter eine Person ist, desto weniger stimmt sie den erwarteten positiven Effekten durch einen KI-Einsatz zu. Mit Blick auf die Aspekte bessere Entscheidungen durch datenbasierte Analysen und den Anstoß technischer Innovationen lässt sich zudem feststellen, dass Männer die Chancen signifikant häufiger sehen als Frauen.

Risiken

Neben den Chancen durch den Einsatz von KI wurden die Befragten ebenfalls zu fünf unterschiedlichen (potenziellen) Risiken befragt. Diese umfassten a) undurchsichtige und nicht nachvollziehbare Entscheidungen, die auf KI-Systemen beruhen, b) eine Gefahr durch Überwachung und Datenmissbrauch, im Rahmen der Auswertung von großen Datenmengen, c) eine Verstärkung von Vorurteilen und Diskriminierung, bspw. bei der Wohnungssuche oder Kreditvergabe, d) einem unverhältnismäßig hohem Energieverbrauch, sowie e) einem Missbrauch der KI zur Erzeugung von Falschinformationen und Desinformation.

Die Zustimmungsraten zu den wahrgenommenen Risiken variieren stärker als bei den Chancen und liegen zwischen 53% und 82% (vgl. Abbildung 4). 58% der Befragten stimmten zu, dass Entscheidungen, die auf KI-Systemen beruhen, undurchsichtig und nicht nachvollziehbar sind (28% stimmten nicht zu, 14% keine Angabe). 79% stimmten zu, dass KI die Gefahr der Überwachung und des Datenmissbrauchs birgt (15 % stimmten nicht zu, 6% keine Angabe). Im Hinblick auf eine mögliche Verstärkung von Vorurteilen und Diskriminierung stimmten 63% zu (21% stimmten nicht zu, 16% keine Angabe). Das Risiko eines unverhältnismäßig hohen Energieverbrauchs sahen lediglich 53% der Befragten (28% stimmten dem potenziellen Risiko nicht zu, 19% keine Angabe). Im Gegensatz dazu stimmte ein hoher Anteil der Befragten, nämlich 82%, zu, dass KI zur Erzeugung von Falschinformationen und Desinformation genutzt werden kann (12% stimmten nicht zu, 6% keine Angabe).

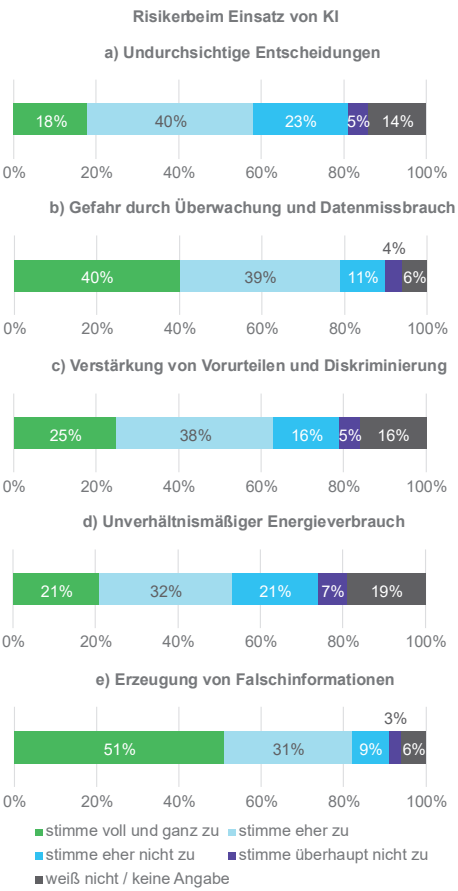


ABBILDUNG 4: RISIKEN BEIM EINSATZ VON KI (QUELLE: EIGENE DARSTELLUNG)

Auch mit Blick auf die sozio-demografischen Variablen lassen sich eine Reihe von Unterschieden feststellen, die in den Tabellen im separaten Anhang ausführlich dargelegt sind. Sowohl für die Gefahren durch Überwachung als auch die Verstärkung von Diskriminierung, den unverhältnismäßigen Energieverbrauch sowie die Erzeugung von Falschinformationen lassen sich signifikant positive Alterseffekte feststellen. Das bedeutet, dass je älter die Befragten sind, desto

mehr stimmten sie den möglichen Risiken zu.[12] Männer stimmten signifikant häufiger als Frauen zu, dass KI die Risiken von undurchsichtigen und nicht nachvollziehbaren Entscheidungen berge, Vorurteile und Diskriminierung verstärke und mit einem unverhältnismäßigen Energieverbrauch einhergehe.[13] Und auch mit Blick auf die digitale Affinität der Befragten, konnte ein signifikanter und negativer Zusammenhang mit der Risikoeinschätzung festgestellt werden. Je digital affiner, desto seltener stimmten die Befragten den Risiken der undurchsichtigen und nicht nachvollziehbaren Entscheidungen, einer Gefahr durch Überwachung und Datenmissbrauch sowie einer Verstärkung von Vorurteilen und Diskriminierung zu.

*5. Erwartungen der Bürger*innen an einen verantwortungsvollen Einsatz von KI*

Vor dem Hintergrund der untersuchten Chancen und Risiken beim Einsatz von KI wurden die Befragten im Anschluss gebeten, ihre Erwartungen an einen verantwortungsvollen Einsatz von KI zum Ausdruck zu bringen. Dabei wurden drei Verantwortungsbereiche bzw. Zuständigkeiten unterschieden, nämlich übergeordnete Erwartungen, Erwartungen an Unternehmen sowie Erwartungen an die Politik.

Übergeordnete Erwartungen

Bei den übergeordneten Erwartungen der Befragten beim verantwortungsvollen Einsatz von KI wurden acht unterschiedliche Aspekte befragt. Diese umfassten a) die Richtigkeit der Informationen, b) die Transparenz und Nachvollziehbarkeit der Entscheidungen, c) klare Verantwortlichkeiten, d) hohe Datenschutzanforderungen, e) faire und diskriminierungsfreie Ergebnisse, f) die Möglichkeit einer menschlichen Überprüfung der Ergebnisse, g) ein energieeffizienter Einsatz der KI, um Umweltbelastungen zu minimieren, sowie h) eine Gestaltung der Systeme, sodass Desinformationen nicht erzeugt werden können. Alle acht Erwartungen erhielten von den Befragten hohe bis sehr hohe Relevanzbewertungen zwischen 79% und 92%, die im Detail der Abbildung 5 entnommen werden können.

---

12 Bei der Gefahr von undurchsichtigen und nicht nachvollziehbaren Entscheidungen war der Effekt nicht signifikant.

13 Bei den Gefahren durch Überwachung und Datenmissbrauch sowie der Erzeugung von Falschinformationen konnte kein Unterschied zwischen den Geschlechtern feststellt werden.

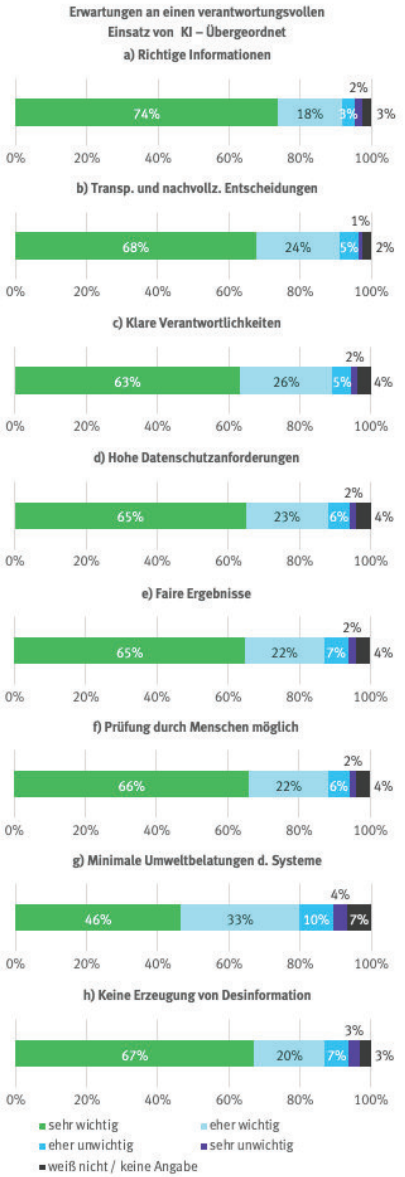


ABBILDUNG 5: ERWARTUNGEN AN EINEN VERANTWORTUNGSVOLLEN EINSATZ VON KI – ÜBERGEORDNET (QUELLE: EIGENE DARSTELLUNG)

92% der Befragten halten die Richtigkeit der Informationen für wichtig (5% unwichtig, 3% keine Angabe).[14] Ebenfalls 92% der Befragten halten transparente und nachvollziehbare Entscheidungen für wichtig (6% unwichtig, 2% keine Angabe). Auch mit Blick auf klare Verantwortlichkeiten zeigen sich hohe Werte in der Wichtigkeit (89%) (7% unwichtig, 4% keine Angabe). Hohe Datenschutzanforderungen haben mit 88% ebenfalls einen hohen Stellenwert (8% unwichtig, 4% keine Angabe). Für 87% der Befragten sind faire und diskriminierungsfreie Ergebnisse wichtig (9% unwichtig, 4% keine Angabe). 88% geben an, dass die Möglichkeit einer menschlichen Überprüfung von Ergebnissen der KI für sie wichtig ist (8% unwichtig, 4% keine Angabe). Den geringsten, jedoch immer noch einen hohen Zustimmungswert hat der energieeffiziente Einsatz im Sinne einer Minimierung von Umweltbelastungen mit 79% (14% unwichtig, 7% keine Angabe). Dass keine Desinformationen durch KI-Systeme erzeugt werden, ist für 87% der Befragten wichtig (10% unwichtig, 3% keine Angabe). Auch im Hinblick auf die sozio-demografischen Attribute der Befragten lassen sich Zusammenhänge bzw. Unterschiede feststellen. Dies trifft zum einen auf das Alter und zum anderen auf das Geschlecht der Befragten zu.[15] Dabei gilt, je älter die Befragten, desto wichtiger bewerten sie die einzelnen Aspekte.[16] Zudem sind die Anforderungen für Frauen signifikant wichtiger als für Männer.

Erwartungen an Unternehmen

In der zweiten Frage zur Einschätzung der Erwartungen wurden die Befragten gebeten, die Wichtigkeit unterschiedlicher Maßnahmen bzw. Anforderungen an Unternehmen zu bewerten. Dabei wurden zehn Punkte unterschieden. Dies waren a) Transparenz, indem Unternehmen den Einsatz von KI offenlegen und die Funktionsweise verständlich erklären, b) eine menschliche Aufsicht, sodass die KI Entscheidungen nicht alleine trifft, c) diskriminierungsfreie Trainingsdaten, d) eine Stärkung der KI-Kompetenzen, sodass Mitarbeiter*innen und Bürger*innen über die Chancen und Risiken der Systeme aufgeklärt werden, e) die Gewährleistung des Datenschutzes, f) eine Einbeziehung von unterschiedlichen Stakeholdern bei der Entwicklung der Systeme, g) das Mo-

14 Wie auch in den vorhergehenden Abschnitten werden die Antwortkategorien zur besseren Lesbar- und Verständlichkeit gepoolt berichtet. Unter „wichtig" fallen sowohl die Antwortkategorie „sehr wichtig" als auch „eher wichtig" und unter „unwichtig" fallen die Antwortkategorien „eher unwichtig" und „sehr unwichtig". Dies trifft auch auf die folgenden Ergebnisse zu den Erwartungen an Unternehmen und Politik zu.

15 Detaillierte Analysen sind im extern-verfügbaren Anhang einsehbar (vgl. CDR-Initiative 2025a).

16 Der Alterseffekt ist mit Blick auf den Wunsch nach minimalen Umweltbelastungen eher schwach, jedoch signifikant.

nitoring und die Evaluation der Systeme, bspw. um Risiken frühzeitig zu erkennen, h) eine Berücksichtigung der Umweltauswirkungen, i) eine Verringerung von Desinformation, indem KI-Inhalte gekennzeichnet werden sowie j) die Gewährleistung von Sicherheit, um vor missbräuchlichen Eingriffen zu schützen. Die Ergebnisse finden sich in Abbildung 6. Insgesamt bewerten die Befragten die unterschiedlichen Anforderungen als wichtig, wobei gewisse Unterschiede existieren, sodass die Relevanzwerte zwischen 75% und 90% liegen.

Die transparente Funktionsweise bewerten 88% der Befragten als wichtig (8% unwichtig, 4% keine Angabe); eine menschliche Aufsicht 87% (8% unwichtig, 5% keine Angabe). Für 83% ist der Einsatz von diskriminierungsfreien Trainingsdaten wichtig (11% unwichtig, 6% keine Angabe). Die Notwendigkeit, die KI-Kompetenz unter Bürger*innen und Mitarbeiter*innen zu stärken, sehen 86% der Befragten als wichtig an (8% unwichtig, 6% keine Angabe). Für 89% ist die Gewährleistung des Datenschutzes wichtig (8% unwichtig, 3% keine Angabe). Ein niedrigerer Wert wird mit Blick auf die Einbeziehung von Stakeholdern erreicht. Diese Anforderung finden 74% der Befragten wichtig und 14% können keine Angabe machen (12% unwichtig). Das Monitoring der KI-Systeme ist für 86% der Befragten wichtig (8% unwichtig, 6% keine Angabe) und die Berücksichtigung der Umweltauswirkungen für 80% (14% unwichtig, 6% keine Angabe. Eine Verringerung von Desinformation ist für 87% wichtig (7% unwichtig, 6% keine Angabe). Zu guter Letzt und zudem mit 90% am höchsten wird die Wichtigkeit bewertet, dass Sicherheit gewährleistet wird (7% unwichtig, 3% keine Angabe). Für 89% ist die Gewährleistung des Datenschutzes wichtig (8% unwichtig, 3% keine Angabe). Ein niedrigerer Wert wird mit Blick auf die Einbeziehung von Stakeholdern erreicht. Diese Anforderung finden 74% der Befragten wichtig und 14% können keine Angabe machen (12% unwichtig). Das Monitoring der KI-Systeme ist für 86% der Befragten wichtig (8% unwichtig, 6% keine Angabe) und die Berücksichtigung der Umweltauswirkungen für 80% (14% unwichtig, 6% keine Angabe. Eine Verringerung von Desinformation ist für 87% wichtig (7% unwichtig, 6% keine Angabe). Zu guter Letzt und zudem mit 90% am höchsten wird die Wichtigkeit bewertet, dass Sicherheit gewährleistet wird (7% unwichtig, 3% keine Angabe). Auch hier finden sich einige systematische Unterschiede zwischen den Altersgruppen und Geschlechtern (vgl. Tabellen im separaten Anhang). z, d. h. je älter die Befragten, desto wichtiger sind die unterschiedlichen Anforderungen.[17] Auch mit Blick auf das Geschlecht bestätigen sich die vorherigen Ergebnisse. So bewerten Frauen die Wichtigkeit der Anforderungen signifikant höher als Männer.

[17] Der Zusammenhang ist schwächer für die Anforderungen diskriminierungsfreie Trainingsdaten und Umwelt. Das heißt, die Unterschiede zwischen den Altersgruppen sind nicht so ausgeprägt wie bei den anderen Anforderungen.

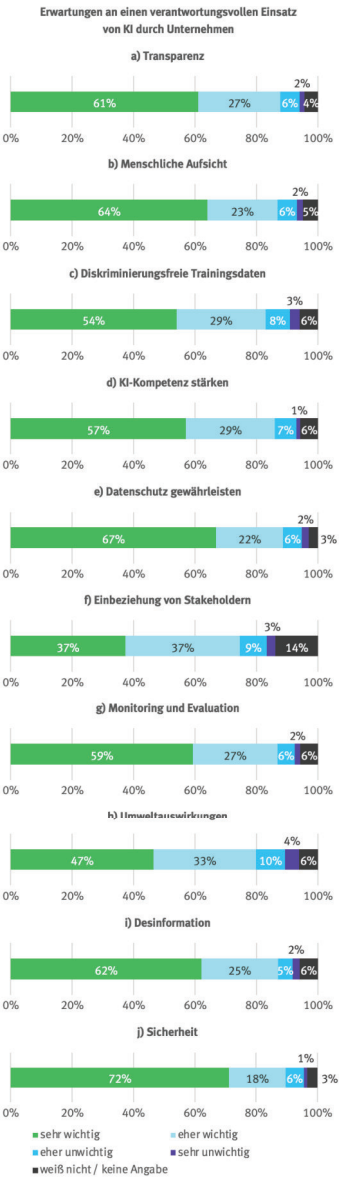


ABBILDUNG 6: ERWARTUNGEN AN EINEN VERANTWORTUNGSVOLLEN EINSATZ VON KI DURCH UNTERNEHMEN (QUELLE: EIGENE DARSTELLUNG)

Erwartungen an die Politik bzw. Regulierung

Zuletzt konnten die Befragten ihre Erwartungen an Politik und Regulierung zum Ausdruck bringen. Hierbei wurden fünf unterschiedliche Anforderungen mit Blick auf ihre Wichtigkeit aus Sicht der Befragten bewertet. Sie umfassen a) einfache Beschwerdemöglichkeiten, d.h. die Politik sollte sicherstellen, dass sich Bürger*innen bei Rechtsverstößen auf einfache Weise bei einer einheitlichen Stelle beschweren können, b) Haftungsregeln, d.h. die Politik sollte eindeutige Haftungsregeln schaffen, die sicherstellen, dass die Verantwortlichen für Schäden oder Fehlentscheidungen zur Rechenschaft gezogen werden, c) eine einheitliche Durchsetzung, d.h. die Politik sollte dafür sorgen, dass die bestehenden Regeln in ganz Europa einheitlich durchgesetzt werden, d) Regeln setzen, d.h. die Politik sollte technologische Entwicklungen genau beobachten und bei Bedarf weitere Regeln für den Einsatz von KI schaffen, und e) die Untersagung von Gesichtserkennung, d.h. der Einsatz von KI zur Gesichtserkennung in der Öffentlichkeit durch Unternehmen, z. B. in Supermärkten, sollte untersagt werden.

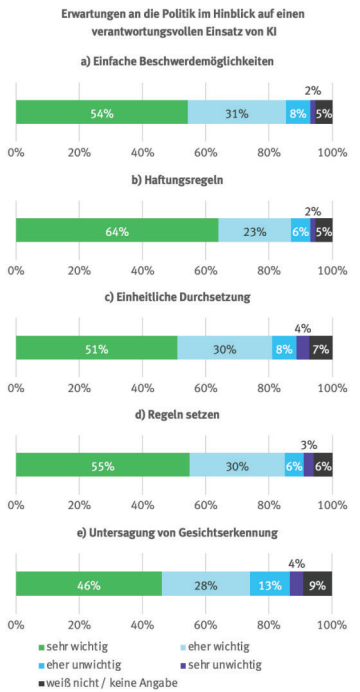


ABBILDUNG 7: ERWARTUNGEN AN DIE POLITIK IM HINBLICK AUF EINEN VERANTWORTUNGSVOLLEN EINSATZ VON KI (QUELLE: EIGENE DARSTELLUNG)

Auch in diesem Fall finden sich in den Ergebnissen hohe Relevanzbewertungen zwischen 74% und 87%, die in Abbildung 7 dargestellt sind.

Für 85% der Befragten sind einfache Beschwerdemöglichkeiten wichtig (10% unwichtig, 5% keine Angabe). Haftungsregeln werden von 87% der Befragten als wichtig erachtet (8% unwichtig, 5% keine Angabe). Die einheitliche Durchsetzung von Regeln ist für 81% der Befragten wichtig (12% unwichtig, 7% keine Angabe). Für 85% sind weitere Regeln in Abhängigkeit des technologischen Fortschritts wichtig (9% unwichtig, 6% keine Angabe). Den geringsten, aber dennoch hohen Wert erhält mit 74% die Untersagung von Gesichtserkennung (17% unwichtig, 9% keine Angabe). Auch in diesem Fall decken sich viele bereits identifizierte Effekte der sozio-demografischen Variablen (vgl. Tabellen im separaten Anhang). Weiterhin stabil ist der beobachtete, signifikante und positive Alterseffekt, d.h. je älter die Befragten, desto wichtiger bewerten sie die Anforderungen.[18] Für die Anforderungen bzgl. Haftungsregeln, einheitlicher Durchsetzung und Regelsetzung kann zudem festgestellt werden, dass Frauen diese signifikant wichtiger bewerten als Männer, während bei der Untersagung von Gesichtserkennung Männer eine höhere Wichtigkeit als Frauen an den Tag legen.[19]

6. Fazit

Die Befragung untersucht die Nutzung von KI-Anwendungen durch Bürger*innen sowie deren Erwartungen an verantwortungsvolle KI. Der Einsatz von KI kann entlang des CDR-Kodex-Prinzips „Fairness" entwickelt und gestaltet und somit Zugang und Teilhabe an der digitalisierten Welt stärken oder zum Gegensatz führen. Die Chancen, die somit KI bietet – wie die Erhöhung des Komforts durch Automatisierung, personalisierte Empfehlungen und datenbasierte Entscheidungsunterstützung – werden von den meisten Befragten positiv bewertet. Gleichzeitig sieht eine deutliche Mehrheit der Bürger*innen ebenfalls Risiken beim

18 Eine Ausnahme bildet dabei die Unterlassung der Gesichtserkennung, bei der der Effekt nur sehr schwach ausgeprägt ist.

19 Kein systematischer Unterschied existiert bei den einfachen Beschwerdemöglichkeiten, die für Männer und Frauen ähnlich wichtig sind. Im Hinblick auf die digitale Affinität finden sich lediglich schwache bis sehr schwache Effekte für die Anforderungen der einfachen Beschwerdemöglichkeiten, Haftungsregeln und der Untersagung von Gesichtserkennung. Diese sind negativ und deuten damit in der Tendenz darauf hin, dass je digital affiner die Befragten sind, desto unwichtiger sind die genannten Anforderungen für sie.

Einsatz von KI, insbesondere bezüglich Überwachung und Datenmissbrauch, der Verstärkung von Vorurteilen und Diskriminierung und der Erzeugung von Falschinformationen.

Die Ergebnisse der Umfrage verdeutlichen zudem, dass Bürger*innen hohe Erwartungen an einen verantwortungsvollen und fairen Einsatz von KI haben. Dazu zählen klare Verantwortlichkeiten und transparente Entscheidungen, faire und diskriminierungsfreie Ergebnisse, die Verringerung von Desinformation, ein Monitoring der KI-Anwendung, die Möglichkeit menschlicher Überprüfung sowie strikte Datenschutzvorgaben. Sowohl Unternehmen als auch die Politik stehen vor der Aufgabe, Maßnahmen umzusetzen und Leitplanken zu schaffen, die diese Erwartungen erfüllen und den sicheren, ethischen und fairen Einsatz von KI gewährleisten. Dazu zählen aus Bürger*innensicht insbesondere auch einfache Beschwerdemöglichkeiten und Haftungsregeln. Außerdem zeigen die Ergebnisse, dass demografische Faktoren die Nutzung und Akzeptanz von KI stark prägen. Jüngere und digital affiniere Personen bewerten die Chancen größer, ältere und weibliche Personen haben hingegen höhere Erwartungen an Unternehmen und Politik, KI verantwortungsvoll zu gestalten. Dies verdeutlicht die Notwendigkeit, KI-Systeme so zu gestalten, dass sie den Erwartungen verschiedener Bevölkerungsgruppen gerecht werden und mögliche Barrieren und Bedenken adressieren.

Literaturverzeichnis

BCG (2024): Zwischen Kenntnis und Nutzung von ChatGPT liegen oft Welten – im privaten Kontext stärker als im Beruf, URL: https://www.bcg.com/press/25april2024-zwischen-kenntnis-und-nutzung-von-chatgpt-liegen-oft-welten-im-privaten-kontext-starker-als-im-beruf (aufgerufen am: 18/06/2025).

CDR-Inititative (2025a): CDR-Umfrage #cdrk24. Umfrage zur Bewertung von Chancen und Risiken beim Einsatz von KI in der Bevölkerung, URL: https://cdr-initiative.de/uploads/files/CDR-Initiative-Umfrage-Chancen-und-Risiken-beim-Einsatz-von-KI-Anhang.pdf (aufgerufen am 18/07/2025).

– (2025b): Der CDR-Kodes, URL: https://cdr-initiative.de/kodex (aufgerufen am: 02/09/25).

Kero, S. / Akyürek, S. Y. / Flaßhoff, F. G. (2023): Bekanntheit und Akzeptanz von ChatGPT in Deutschland. Factsheet Nr. 10 – Dezember 2023. Meinungsmonitor, Künstliche Intelligenz, URL: https://www.cais-research.de/wp-content/uploads/Factsheet-10-ChatGPT.pdf (aufgerufen am: 18/06/2025).

Von Datenknappheit zur Datenhoheit[1]

Endkonsument*innen als aktive Gestalter*innen für mehr Fairness in der Datensammlung

K. Valerie Carl

1. Einleitung

Daten spielen für die Entwicklung Künstlicher Intelligenz (KI)-basierter Systeme eine zentrale Rolle. Insbesondere im Kontext fairer KI wird die Bedeutung von Daten zunehmend diskutiert, da Daten eine zentrale Quelle für Unfairness darstellen können (z. B. Feuerriegel et al. 2020). Dabei spielt beispielsweise das Set an Trainings- und Testdaten eine große Rolle, um möglichst diskriminierungsfreie Systeme zu entwickeln. Auch die Art des Trainings, die Wahl der Modelle und wie diese Systeme implementiert werden, sowie die Wahl entsprechender Algorithmen (z. B. Wahl von Pre- oder Post-Processing Algorithmen) werden im Sinne fairer KI vielfach diskutiert (z. B. von Zahn et al. 2022).

1 Am Projekt COMET arbeiten das Deutsche Forschungszentrum für Künstliche Intelligenz (DFKI) und die Goethe Universität Frankfurt/Main (GU) zusammen. An der Entwicklung der Mechanismen haben Prof. Dr. Oliver Hinz (GU), Prof. Dr. Oliver Thomas (DFKI), Constantin Brîncoveanu (GU), Dr. K. Valerie Carl (GU), Léon Dankert (DFKI), Dr. Tobias Dreesbach (DFKI), Rebecca Heigl (GU) und Aaron Witzki (GU) zusammengearbeitet. Das Projekt COMET wird vom Bundesministerium für Wirtschaft und Energie (BMWE) im Rahmen des Technologieprogramms „SmartLivingNEXT – Künstliche Intelligenz für nachhaltige Lebens- und Wohnumgebungen" gefördert. Mit SmartLivingNEXT entsteht ein universelles, KI-basiertes Ökosystem für die einfache und kostengünstige Entwicklung intelligenter und an Nachhaltigkeit ausgerichteter Smart Living-Dienste und -Anwendungen. Insbesondere der unkomplizierte und sichere Datenaustausch zwischen allen Beteiligten wird damit ermöglicht.

Fairness kann in Bezug auf Daten noch weitergedacht werden. So kann faire KI bereits einen Schritt davor beginnen und auch eine faire Akquise von Daten von Endkonsument*innen, also den Datenerzeuger*innen, beinhalten (siehe Abbildung 1). Dabei gehen Gedanken hinsichtlich Fairness über diverse und diskriminierungsfreie Daten hinaus und beinhalten auch die Art und Weise wie Daten für das Training KI-basierter Systeme gesammelt werden. Das heißt beispielsweise, dass sich Endkonsument*innen tatsächlich bewusst sind, dass gerade Daten von ihnen gesammelt werden, an wen diese weitergegeben werden und wofür (bspw. für das Training KI-basierter Systeme) diese Daten eingesetzt werden. Oftmals ist dies in der Praxis nicht der Fall (vgl. Acquisti et al. 2018).

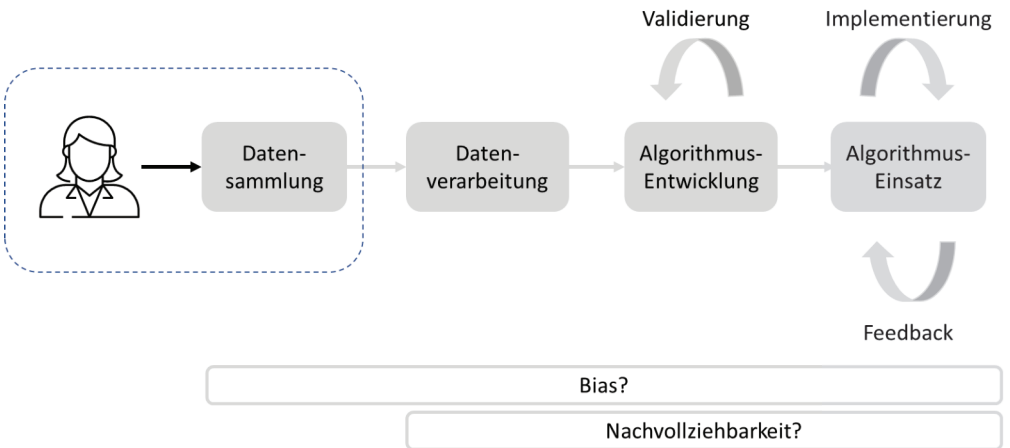


ABBILDUNG 1: BEISPIELHAFTE ENTSCHEIDUNGEN FÜR FAIRE KI (QUELLE: EIGENE DARSTELLUNG IN ANLEHUNG AN VAN HOUTEN 2020)

Endkonsument*innen-Daten werden nicht nur in den Unternehmen genutzt, mit denen Endkonsument*innen eine direkte Interaktion haben, sondern teilweise auch über Unternehmensgrenzen weitergegeben. So werden sie beispielsweise zwischen Töchtergesellschaften und der Muttergesellschaft geteilt. Aber auch unabhängige Unternehmen können ein Interesse an diesen Endkonsument*innen-Daten haben. Insbesondere kleine und mittelständische Unternehmen (KMUs) und Startups stehen oftmals vor einem Henne-Ei-Problem: Für das Training KI-basierter Systeme werden zumeist (größere) Datenmengen benötigt und stellen damit eine Voraussetzung für das

Angebot innovativer digitaler Dienste und Produkte dar. Gleichzeitig sind derartige Daten vor Angebot solcher Produkte und Dienstleistungen in KMUs und Startups häufig nicht vorhanden (vgl. Witzki et al. 2025a). Entsprechend stellt sich die Frage, wie Daten für KI-basierte Services (extern) akquiriert werden können. Oftmals werden diese nicht über Unternehmensgrenzen hinweg zur Verfügung gestellt. Dafür können Datenschutzbedenken, Wettbewerbsvorteile, Interoperabilitätsprobleme oder Vertrauen in fremde Organisationen Gründe sein. Zudem ist die Weitergabe von Daten über Unternehmensgrenzen hinweg auch aus Fairness-Perspektive kritisch zu evaluieren. Ohne eine ausreichende Datengrundlage ist allerdings das Training KI-basierter Systeme nicht möglich. Gerade im Sinne eines fairen Wettbewerbs und Chancen für KMUs und Startups sowie das Aufbrechen sogenannter Datensilos (vgl. Tallon 2013) ist die Verfügbarkeit von Daten über Unternehmensgrenzen hinweg in einer fairen Art und Weise ein Schlüssel für innovative, verantwortungsvolle Systeme der Zukunft, und damit die Verfügbarkeit von Daten außerhalb einiger weniger Unternehmenskonglomerate.

Digitale Ökosysteme, speziell Datenökosysteme, bieten eine sich stetig entwickelnde Plattform, Daten zu akquirieren (vgl. Oliveira et al. 2019; Scheider et al. 2023). Sie können einen rechtlichen sowie Werte-Rahmen für den Austausch von Daten bieten sowie spezifische Mechanismen bereitstellen, die den Datenaustausch über Unternehmensgrenzen hinweg gesichert ermöglichen. Derartige Datenökosysteme entwickeln sich zunehmend in der Realität, beispielsweise in den Domänen Smart Living und Mobilität.

Bisher teilen im Rahmen von Datenökosystemen insbesondere Unternehmen untereinander Daten, die oftmals von Endkonsument*innen erzeugt werden (vgl. Scheider et al. 2023). Im Sinne von mehr Fairness, ist die Einbindung von Endkonsument*innen als aktive Akteure in diesen Austausch möglich. Für eine faire Akquise können in diesen Ökosystemen zusätzliche Mechanismen implementiert werden, die Endkonsument*innen ermächtigen, aktiv zu handeln und damit Daten fair und informiert bereitzustellen (vgl. Dreesbach et al. 2025). So behalten Endkonsument*innen die Hoheit über ihre eigenen Daten, werden in ihrer Souveränität gestärkt und gleichzeitig werden so gegebenenfalls mehr Daten über Unternehmensgrenzen hinweg fair und verantwortungsvoll verfügbar.

Im Folgenden werden zunächst Faktoren und Einflüsse auf eine faire Datensammlung über Unternehmensgrenzen hinweg im Kontext KI-basierte Systeme aufgezeigt, um ein Verständnis

zu ermöglichen, welche Aspekte zu einer fairen Datensammlung für KI-basierte Systeme beitragen können. Darauf aufbauend werden spezifische Mechanismen sowie Design-Entscheidungen in einem digitalen Datenökosystemen beleuchtet, die eine solche faire Datenbereitstellung in der Praxis unterstützen und flächendeckend möglich machen können. So wird für die Domäne Smart Living exemplarisch gezeigt, wie eine fairere Datensammlung sichergestellt werden kann. Die entwickelten Mechanismen sind domänenunabhängig anwendbar und sollen exemplarisch notwendige Entscheidungen und mögliche Designvorschläge aufzeigen. Zuletzt wird ein Ausblick auf mögliche weitere Entwicklungen gegeben.

2. Faire Datensammlung für KI-basierte Systeme

Im Sinne fairer KI und der freiwilligen Übernahme zusätzlicher Verantwortung, können Organisationen über gesetzliche Anforderungen hinausdenken und stärker im Sinne fairer KI handeln, als sie verpflichtet sind (vgl. Mihale-Wilson et al. 2021). Soll also die Datensammlung für KI-basierte Systeme besonders fair und verantwortungsvoll ausgestaltet werden, bieten rechtlich bindende Anforderungen einen guten Ausgangspunkt. Die Deutsche Datenschutz-Grundverordnung (DSGVO) hat für informiertes Einverständnis von Endkonsument*innen den rechtlichen Grundstein gelegt. In der Praxis geben Endkonsument*innen jedoch zumeist zu (längeren, teilweise sehr komplizierten) Datenschutzerklärungen per Checkmark die Zustimmung – oftmals ohne diese wirklich zu lesen.[2] Damit erfolgt zwar eine informierte Einwilligung, die Frage ist allerdings, wie bewusst und fair diese Zustimmung ist und ob Endkonsument*innen immer die Tragweite ihrer Entscheidungen bewusst ist. Unternehmen können zudem Interfaces so gestalten, dass Endkonsument*innen zur Zustimmung zusätzlich incentiviert („nudging“) werden (vgl. Acquisti et al. 2018). Insbesondere im Fall digitaler (Daten-)Ökosysteme erfolgt diese Zustimmung gegebenenfalls nicht nur für das Unternehmen, mit dem direkt eine Geschäftsbeziehung besteht, sondern im Rahmen der Datenschutzerklärung wird gegebenenfalls der Weitergabe an Unternehmen im größeren Unternehmensverband oder sogar für das ganze Ökosystem zugestimmt. Im Sinne einer möglichst fairen Datensammlung ist also die Frage, ob die Zustimmung zu regulären

[2] Einige Unternehmen ergreifen hier bereits zusätzliche Maßnahmen, um eine informierte Einwilligung zu unterstützen (z. B. One-Pager, FAQs, einfache Erläuterungen). Jedoch sind solche zusätzlichen Informationen und unterstützenden Angebote noch nicht flächendeckend verfügbar.

Datenschutzerklärungen (und AGBs) wirklich die volle Souveränität von Endkonsument*innen gewährleistet und eine faire Datenweitergabe über Unternehmensgrenzen hinweg sicherstellt.

Selbst wenn diese Erklärungen gelesen werden, ist die Frage, ob Endkonsument*innen immer verstehen, wie und wofür ihre Daten an wen weitergegeben werden. Aufgrund der teils komplexen und juristischen Formulierungen kann diese für Endkonsument*innen teilweise schwer nachvollziehbar sein. Entsprechend ist im Sinne einer fairen Datensammlung fraglich, ob in jedem Falle eine tatsächlich informierte Entscheidung getroffen werden kann. Im Sinne einer möglichst fairen Datenverwendung sollte Endkonsument*innen klar und verständlich erläutert werden, wofür ihre Daten durch wen genutzt werden (vgl. Carl/Hinz 2024), um eine tatsächlich informierte Zustimmung zu gewährleisten.

Zudem können oftmals keine Anpassungen der Datenweitergabe vorgenommen werden. Vielmehr müssen Endkonsument*innen, sofern sie ein bestimmtes Produkt oder einen bestimmten Service nutzen wollen, der Datenschutzerklärung und damit der hierin dargelegten Datenverwendung zustimmen. Oftmals können nicht für einzelne Zwecke Anpassungen vorgenommen werden. Sind solche Einstellungen anpassbar, steuern trotz allem manche Unternehmen die Datennutzung über die sogenannten Default-Settings, also Voreinstellungen, die festgelegt werden und die mit zusätzlichem Aufwand angepasst werden müssen. Endkonsument*innen müssen in diesem Fall aktiv Änderungen vornehmen, was von vielen Endkonsument*innen nicht genutzt wird (vgl. Acquisti et al. 2018). Eine Personalisierung ist nur durch Initiative von Endkonsument*innen möglich, oftmals nicht bekannt oder wird nicht genutzt. Zudem steuern Unternehmen über den damit verbundenen (zeitlichen) Aufwand, wie viele Anpassungen tatsächlich vorgenommen werden. Je länger eine entsprechende Anpassung dauert, desto weniger wahrscheinlich ist, dass viele Endkonsument*innen ihre Möglichkeit der Anpassung nutzen.

Endkonsument*innen haben in der Praxis oftmals nicht das Gefühl, die Datenverarbeitung tatsächlich kontrollieren zu können (vgl. Scheider et al. 2023). Diese Kontrolle ist für viele Endkonsument*innen allerdings eine zentrale Voraussetzung für das Teilen von Daten (vgl. Aitken et al. 2016). Zudem ist die Frage, wie einfach diese Kontrolle ausgeübt werden kann (vgl. Hanneke et al. 2023). Je niederschwelliger Lösungen sind, desto eher werden Endkonsument*innen ihre Rechte nutzen.

Datenakquise und deren Verwendung für KI ist nicht per se unverantwortlich, vielmehr können Maßnahmen ergriffen und Mechanismen implementiert werden, die diese fairer und damit

verantwortungsvoller machen. Wichtig ist das Bewusstsein von Organisationen, dass faire KI bereits bei der Akquise von Daten für das Training beginnt. Oftmals ist eine Fremdakquise über Unternehmensgrenzen hinweg notwendig, weil entsprechende Daten im Unternehmen nicht verfügbar sind. Dabei ist ebenfalls zu beachten, dass Verantwortung nicht an der eigenen Unternehmensgrenze aufhört, sondern das akquirierende Unternehmen auch bewusste Entscheidungen hinsichtlich aller beteiligten Akteure in Betracht ziehen sollte. So sollte sich das akquirierende Unternehmen fragen, ob die Daten von Endkonsument*innen als Datenerzeuger*innen verantwortungsvoll erhoben wurden und eine verantwortungsvolle und faire Weitergabe stattfindet. Somit ergeben sich im Sinne einer fairen Datenakquise und Verwendung als Baustein fairer KI folgende Schlüsselfragen:[3]

- SF1: Sind sich Endkonsument*innen der Datenverwendung bewusst?
- SF2: Erfolgt die Datenverwendung für Endkonsument*innen unerwartet über Unternehmensgrenzen hinweg?
- SF3: Ist die Datenverwendung für Endkonsument*innen verständlich?
- SF4: Kann eine Personalisierung der Datenverwendung durch Endkonsument*innen vorgenommen werden?
- SF5: Haben Endkonsument*innen die Kontrolle über das Teilen der Daten?
- SF6: Erfolgt die Datensammlung, -akquise und -verwendung bei allen beteiligten Akteuren verantwortungsvoll?

3. Datenökosysteme als Treiber für faire Datenakquise bei gleichzeitig höherer Verfügbarkeit

Digitale Datenökosysteme stellen eine Plattform für den Austausch von Daten dar (vgl. Oliveira et al. 2019). Je nach Governance und Struktur kann ein solcher Datenaustausch mehr oder minder verantwortungsvoll vonstattengehen. Bisher werden dabei vor allem zwei Akteure betrachtet: organisationale Datenbereitsteller (z. B. App-Anbieter) und organisationale Datenkonsumenten (z. B. KI-Service-Anbieter) (siehe Abbildung 2) (vgl. Scheider et al. 2023). Zukünftig

[3] Die vorgestellten Schlüsselfragen sollen als Diskussionsanstoß dienen und stellen keine abschließende Aufzählung dar. Auf Basis bisheriger Forschung und Beobachtungen in der Praxis wurden besonders präsente Schlüsselfragen ausgewählt, die eine zentrale Rolle für faire Datenakquise spielen.

könnten Endkonsument*innen allerdings aktiver in diesen Datenaustausch eingebunden werden und souverän handeln (siehe Abbildung 2) (vgl. Oliveira et al. 2019).

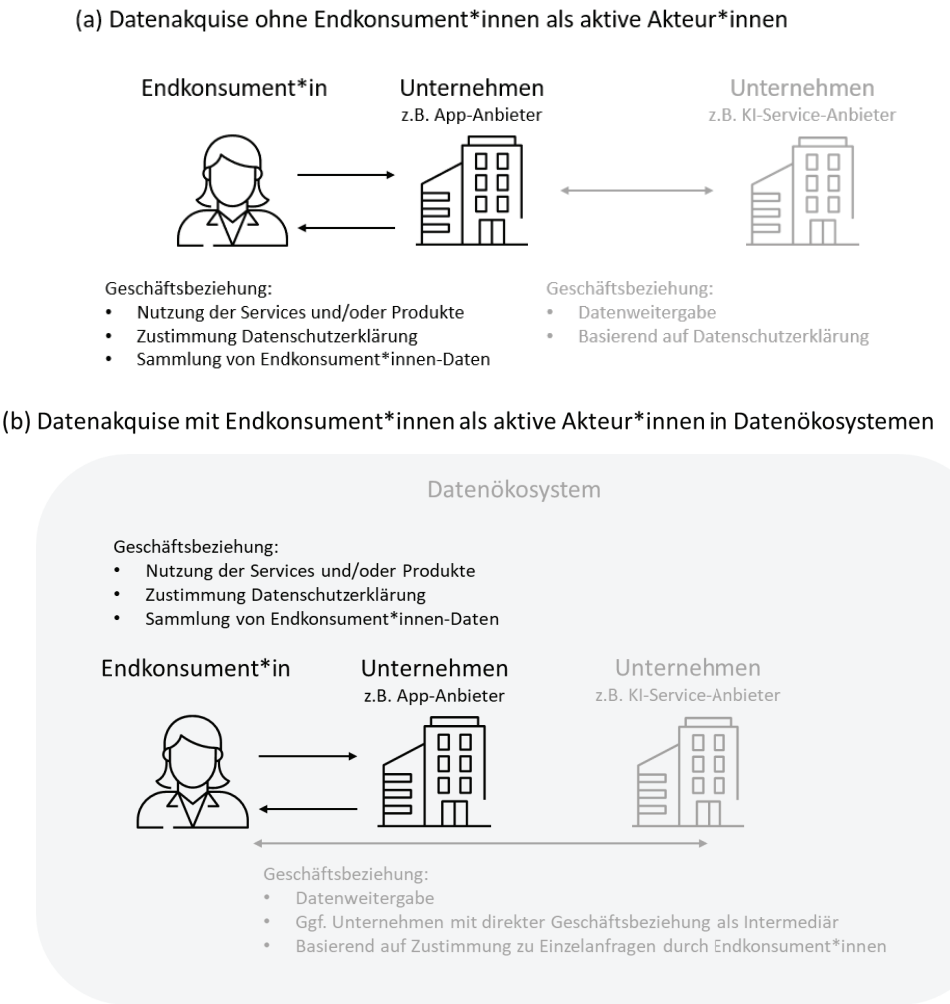


ABBILDUNG 2: DATENAKQUISE VON ENDKONSUMENT*INNEN-DATEN (QUELLE: EIGENE DARSTELLUNG)

Wenn Endkonsument*innen als souveräne Entitäten in Datenökosystemen agieren und ihre Daten informiert und bewusst für organisationale Datenkonsumenten bereitstellen können, werden einerseits Endkonsument*innen in ihren Rechten gestärkt. Dafür sollten sie ihre Rechte aber auch möglichst einfach ausüben können, um die Hürde für ihre Ausübung möglichst gering zu halten. So wird faire KI über Training und Implementierung hinausgedacht, indem auch eine faire Datenakquise verfolgt wird. So kann gegebenenfalls auch die Datenverfügbarkeit über Unternehmensgrenzen hinweg unterstützt werden. Wenn Endkonsument*innen ihre Daten freiwillig und informiert an andere Unternehmen übertragen, fallen für das bisher verwaltende Unternehmen (z. B. organisationaler App-Anbieter) rechtliche Bedenken weg, die diese Übertragung bisher behindert haben. Zudem können Endkonsument*innen zusätzlich incentiviert sein, ihre Daten freiwillig und informiert zur Verfügung zu stellen. Ähnlich wie beim Crowdfunding können Daten für das Training innovativer KI-basierter Services bereitgestellt werden mit der Motivation, diesen Service zukünftig selbst nutzen zu können. Beispielsweise in der Medizinforschung ist dieser Effekt bereits zu beobachten. Patient*innen stellen ihre Daten freiwillig zur Verfügung, damit eigene Krankheiten besser erforscht werden können oder um der Gesellschaft als Ganzes zu helfen (z. B. Aitken et al. 2016). Die meisten KI-basierten Systeme fallen nicht in die Kategorie der Medizinforschung. Trotz allem kann beispielsweise die Aussicht auf bessere Energieprognosen und Kosteneinsparpotenziale oder mehr Komfort Endkonsument*innen motivieren, ihre Daten freiwillig für das Training von KI-basierten Systemen bereitzustellen.

Um diese Souveränität zu ermöglichen und eine fairere Datenakquise zu verwirklichen, müssen verschiedene Mechanismen und Governances in digitalen Datenökosystemen implementiert werden, die Endkonsument*innen die direkte und einfache Kontrolle über ihre Daten geben und gleichzeitig die Möglichkeit, ihre Daten souverän freizugeben.

In der Domäne Smart Living werden solche Mechanismen im SmartLivingNEXT Ökosystem[4] durch das Projekt COMET[5] pilotiert (siehe Abbildung 3). Diese Mechanismen stellen einen möglichen Weg dar, Endkonsument*innen zu souveränen Akteuren in digitalen (Daten-)Ökosystemen zu ermächtigen und ein bewusstes und faires Datenteilen zu unterstützen. Endkonsument*innen werden über bestehende Apps aus dem Bereich Smart Living (z. B. Energieverbrauchsmonitoring und Assisted Living) an das digitale Ökosystem angebunden. Die pilotierten Mechanismen sind sowohl in

4 Für eine ausführliche Einführung des Ökosystems vgl. SmartLivingNext (2025).

5 Für eine ausführliche Einführung des Projektes vgl. SmartLivingNext in COMET (2025).

bestehenden Apps anwendbar wie im Projekt gezeigt wird, als auch in eigens entwickelten Apps umsetzbar. So können bestehende Unternehmen und Datenökosysteme aber auch sich neu entwickelnde Ökosysteme und Anwendungen auf die pilotierten Mechanismen zurückgreifen und sie domänenunabhängig implementieren. Anhand des Projektes COMET und der Domäne Smart Living soll exemplarisch gezeigt werden, wie Datensouveränität bei Endkonsument*innen einerseits zu faireren KI-basierten Systemen und andererseits zu besserer Datenverfügbarkeit führen kann.

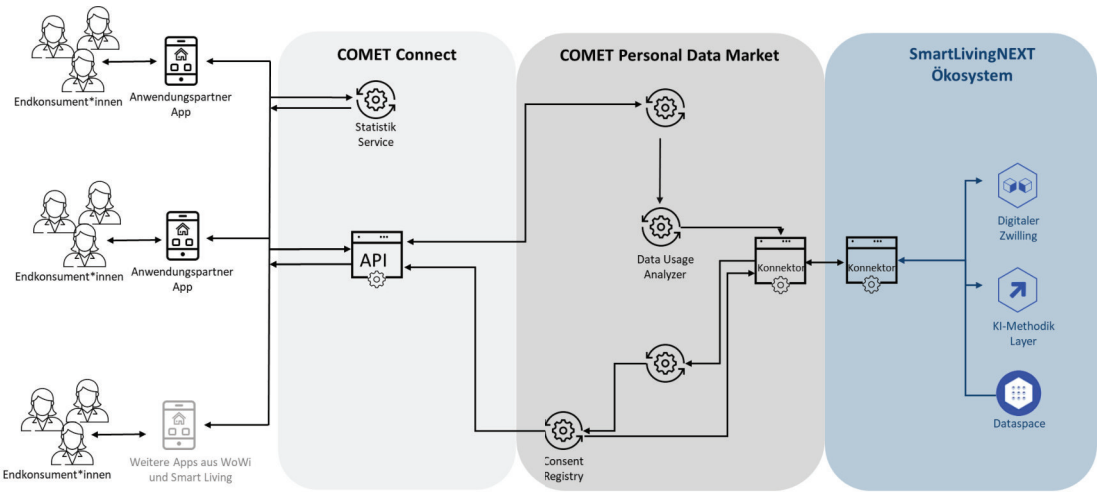


ABBILDUNG 3: PILOTIERTE MECHANISMEN IM PROJEKT COMET (QUELLE: EIGENE DARSTELLUNG IN ANLEHUNG AN COMET 2025)

Die genannten sechs Schlüsselfragen (siehe Kapitel 2) waren zentral bei der Entwicklung entsprechender Mechanismen im Projekt COMET. Die Sicherstellung der fairen Datensammlung über verschiedene Akteure hinweg (SF6) wird primär durch die aktive Einbindung von Endkonsument*innen als souveräne Entitäten adressiert. Endkonsument*innen sollen im Rahmen dieser Mechanismen selbst entscheiden können, wer ihre Daten erhält und was mit ihnen passiert und treten damit an die Stelle der bisher häufigen Entscheidungsträger: organisationale Datenbereitsteller wie App-Anbieter. Diese Souveränität soll möglichst niederschwellig und einfach ausübbar sein. Endkonsument*innen erhalten eine aktive und selbstbestimmte Rolle im digitalen Datenökosystem, sodass akquirierende Unternehmen direkt mit Endkonsument*innen in Kontakt kommen und nicht wie bisher zumeist mit einem

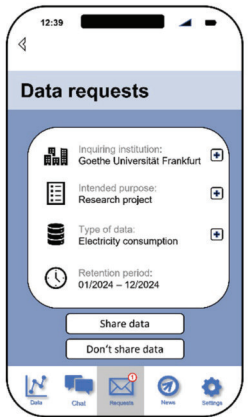
zwischengeschalteten Unternehmen, mit dem die Datenerzeuger*innen in direkter Interaktion stehen. Durch diese direkte Interaktion mit Endkonsument*innen kann von einer faireren Datenakquise ausgegangen werden, da Endkonsument*innen ihre Daten bewusst, informiert, fair und freiwillig teilen. Entsprechend übernimmt das akquirierende Unternehmen auch über Unternehmensgrenzen hinweg Verantwortung, indem eine faire Datensammlung Voraussetzung für die Akquise der Daten ist. Es werden also nur Daten akquiriert, die fair, informiert und freiwillig gewonnen wurden durch die direkte Einbindung von Endkonsument*innen. Im Gegensatz dazu, sind beispielsweise unbewusst erhobene Scrolling-Daten weniger verantwortungsvoll erhobene Daten, worüber sich auch das akquirierende Unternehmen bewusst sein und dafür Verantwortung übernehmen sollte.

Darüber hinaus sollen in diesem exemplarischen Ökosystem Endkonsument*innen für die freiwillige und faire Datenfreigabe einen persönlichen Vorteil erhalten, unabhängig davon, wie oft oder regelmäßig sie ihre Daten freigeben, um die tatsächliche Freiwilligkeit der Freigabe sicherzustellen. Derart freigegebene Daten sind Innovations-Enabler, sodass Endkonsument*innen von innovativeren Produkten und Services profitieren. Darüber hinaus wird ihnen von einer zentralen Entität des digitalen Ökosystems SmartLivingNEXT ein weiterer Service zur Verfügung gestellt, der illustriert, was mit freigegeben Daten erreicht werden kann. Im Falle des SmartLivingNEXT Ökosystems und des Projektes COMET ist dies ein sogenannter „Statistik-Service", der Energie- und Assisted Living-Daten auswertet und mit Hilfe von KI personalisierte Empfehlungen, Vorhersagen und Warnungen (in den Bereichen Energie und Assisted Living) ausgibt. So sollen Endkonsument*innen durch das Ökosystem einen persönlichen Vorteil erhalten als Motivation für mögliche freiwillige Datenfreigaben. Gleichzeitig wird mit diesem Statistik-Service gezeigt, welche Systeme mit Hilfe derartiger Datenfreigaben zukünftig durch die akquirierenden Unternehmen entwickelt werden könnten. Damit geht das digitale Ökosystem in Vorleistung gegenüber den Endkonsument*innen ohne erwartete Gegenleistung. Wichtig dafür ist, dass Datenfreigaben keine Voraussetzung für die Nutzung dieses Service sind. Vielmehr soll der Service illustrieren, was mit Hilfe von Datenfreigaben möglich ist und einen Mehrwert für das digitalen Datenökosystem als Ganzes darstellen. In verschiedenen Use Cases in der Domäne Smart Living konnte bereits gezeigt werden, dass durch derartige Statistik-Services ein Mehrwert für Endkonsument*innen entstehen kann (z. B. Brîncoveanu et al. 2024).

Ein weiterer zentraler pilotierter Mechanismus ist die sogenannte „Consent Registry". Dieser Mechanismen hat Einzelanfragen für die Datenfreigabe zum Zweck, die von Endkonsument*innen im Sinne einer gestärkten Souveränität geschätzt werden (vgl. Hanneke et al. 2023). Das heißt

Endkonsument*innen sollen nicht durch (einmalige) Zustimmung zu einer Datenschutzerklärung ihre Daten freigeben. Vielmehr sollen sie je Verwendungszweck und akquirierendem Unternehmen entscheiden können, ob sie ihre Daten für diesen Zweck freigeben wollen (vgl. Witzki et al. 2025a) (siehe Abbildung 4).

(a) Initiale Ansicht einer Datenanfrage für Endkonsument*innen



(b) Detaillierte Ansicht einer Datenanfrage für Endkonsument*innen (ausklappbare Felder)

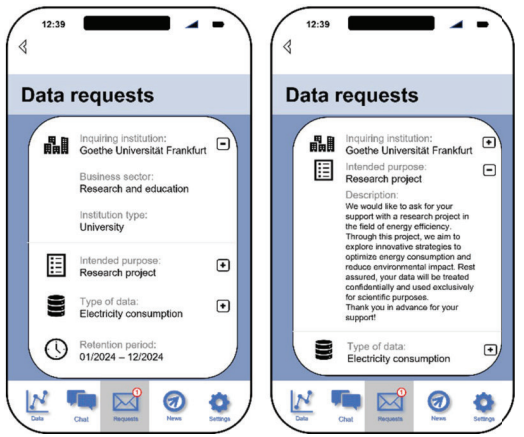


ABBILDUNG 4: ANSICHT VON DATENANFRAGEN FÜR ENDKONSUMENT*INNEN IM RAHMEN DER CONSENT REGISTRY (QUELLE: WITZKI ET AL. 2025A)

Die Consent Registry soll die Souveränität von Endkonsument*innen stärken und eine Personalisierung der Datenfreigaben soll einfach und intuitiv möglich sein (SF4). Durch diese Einzelanfragen soll sichergestellt werden, dass jeder Freigabe bewusst und informiert zugestimmt wird (SF1) und somit der Grundstein für faire Datenweitergabe und damit faire KI gelegt wird. Durch die Verwendung der Consent Registry für Datenweitergabe und die explizit benötigte Zustimmung zu jeglicher Verwendung, entsteht bei Endkonsument*innen ein größeres Bewusstsein und Verständnis von wem und wofür ihre Daten verwendet werden (SF1). So werden Endkonsument*innen in ihrer Datenhoheit und Souveränität gestärkt.

Durch die proaktive Bereitstellung eines Mechanismus zur Datenfreigabe auf Basis einer Opt-in-Lösung, kann eine vollständige Personalisierung der Datenfreigabe erfolgen. Die Datenfreigabe ist keine Voraussetzung für die Nutzung von Diensten oder Produkten, sodass kein Zwang besteht, der Verarbeitung zuzustimmen. Durch eine möglichst einfache und intuitive Einbindung in neue und bestehende Applikationen soll zudem die Nutzungsschwelle so gering wie möglich gehalten werden, um eine Personalisierung möglichst einfach zu machen. So bleibt auch die Kontrolle der Datenfreigabe komplett bei Endkonsument*innen. Die Datenfreigabe erfolgt bewusst und informiert ohne Verpflichtung oder als Trade-off gegenüber der Nutzung von Produkten und Services und somit freiwillig. Dadurch erhalten Endkonsument*innen volle Souveränität und behalten die Datenhoheit (SF5).

Insbesondere die Datenweitergabe über Unternehmensgrenzen hinweg (SF2) wird für Endkonsument*innen dadurch transparenter und eine souveräne Entscheidung kann je Einzelfall getroffen werden. Das heißt, bei gleichem Zweck, Dauer, etc. kann trotzdem zwischen Unternehmen unterschiedlich entschieden werden, weil beispielsweise das Vertrauen zu Unternehmen A größer ist als zu Unternehmen B oder die erwartete Technologie interessanter. Auch im Zeitverlauf kann zwischen gleichen Zweck, Datentyp und Unternehmen abweichend entschieden werden. Es erfolgt eine Zustimmung auf Einzel-Case-Basis, die nicht auf zukünftige Anfragen angewendet wird. Insbesondere wird nicht auf Unternehmensbasis, sondern auf Use-Case-Basis eine Entscheidung zur Datenfreigabe getroffen, sodass kein Unternehmen Blanko-Daten bekommt, mit denen beliebige Zwecke verfolgt werden können (z. B. das Training mehrere verschiedener KI-basierter Services). Unternehmen müssen für jeden geplanten Dateneinsatz eine eigene Anfrage an die Endkonsument*innen stellen. Insgesamt wird so mehr Bewusstsein und Transparenz sichergestellt, was schlussendlich zu einer faireren und bewussteren Datenbereitstellung beiträgt.

Voraussetzung für Anfragen von akquirierenden Unternehmen über die Consent Registry ist zudem, dass Anfragen leicht verständlich formuliert werden. Zudem wird den akquirierenden Unternehmen ein Formular bereitgestellt, das genutzt werden muss, sodass Endkonsument*innen Anfragen immer in der gleichen Struktur erhalten und alle benötigten Informationen für eine informierte Entscheidung bereitgestellt werden (SF3). So ist eine einfachere Vergleichbarkeit zwischen Anfragen gegeben. Eine gleichbleibende Struktur reduziert zudem für Endkonsument*innen den nötigen Aufwand, sich in neue Anfragen einzuarbeiten und entsprechende Entscheidungen zu treffen.

Des Weiteren sollen Endkonsument*innen nicht nur souverän ihre Daten freigeben können, sondern ihre Souveränität und Datenhoheit soll auch langfristig geschützt werden, mit Hilfe des sogenannten „Data Usage Analyzer". Dieser Mechanismus soll Endkonsument*innen die langfristige Nachverfolgbarkeit, was mir ihren Daten geschieht, ermöglichen und ihnen somit die Kontrolle über ihre Daten vereinfachen (SF5). So soll darüber jede getätigte Freigabe ersichtlich sein, wofür diese Daten im akquirierenden Unternehmen genutzt wurden, ob diese bereits gelöscht wurden oder wann diese gelöscht werden und weitere Informationen zur Freigabe langfristig und zugänglich gespeichert werden. So müssen nicht wie nach DSGVO möglich Anfragen zur Datennutzung gestellt werden. Vielmehr sind diese Informationen jederzeit über die angebundenen Apps abrufbar. So wird für Endkonsument*innen die Ausübung ihres Rechts auf Information, Anpassung und Löschung vereinfacht durch einen konsolidierten, einfach nachvollziehbaren Überblick über die Datenverwendung und so ihre Souveränität gestärkt. Damit wird auch nach Datenfreigabe zur Personalisierbarkeit beigetragen (SF4), indem eine persönliche Übersicht der genutzten Daten bereitgestellt wird. Entsprechend hört Verantwortung für implementierende Unternehmen nicht bei der Datenfreigabe selbst auf, sondern auch danach wird weiter Verantwortung im Sinne fairer KI übernommen. Derartige nachträgliche Transparenz über die tatsächliche Verwendung der Daten kann das Vertrauen von Endkonsument*innen stärken und so schlussendlich gegebenenfalls mehr Datenfreigaben begünstigen, da die Transparenz hinsichtlich der Verwendung ermöglicht wird und so die Kontrolle über die Daten bei den Endkonsument*innen unterstützt und leichter ausübbar ist.

Die pilotierten Mechanismen können sowohl innerhalb eines Datenökosystems als auch innerhalb eines einzelnen Unternehmens eingesetzt werden. Unternehmen können auch ihren eigenen Kund*innen Souveränität geben, einfach, verständlich und fair ihre Daten zu verwalten und

Datenhoheit über eigens erzeugte Daten zu behalten. So kann beispielsweise auch an eigene Kund*innen (nicht nur an unternehmensexterne Endkonsument*innen) eine Anfrage gestellt werden, ob Daten für das Training eines bestimmten KI-basierten Systems genutzt werden können. So können Unternehmen die in COMET pilotierten Mechanismen auch für den internen Einsatz weiterdenken und so eine fairere Datenakquise intern verfolgen.

4. Ausblick

Zusammengefasst sollte ein möglicher Ansatz für faire Datenakquise gezeigt und damit ein Beitrag zu fairer KI und der Übernahme von Verantwortung im Sinne von Corporate Digital Responsibility (CDR) geleistet werden. Faire KI umfasst eine Vielzahl verschiedener Bausteine wie diverse und diskriminierungsfreie Trainingsdaten, faire Systeme, bestenfalls erklärbare Entscheidungen und ein ethischer und fairer Einsatz solcher Systeme.

Die hier vorgestellten pilotierten Mechanismen und Anwendungen sollen einen Beitrag dazu leisten, faire KI auch aus einer anderen, weniger diskutierten Brille zu beleuchten: der Fairness in der Datenakquise und -nutzung für das Training von KI. Insbesondere aufgrund des Bedarfs, Daten über Unternehmensgrenzen hinweg verfügbar zu machen, können Endkonsument*innen als souveräne Akteure in digitalen Datenökosystemen eine wichtige Rolle spielen. Einerseits können Endkonsument*innen so souveräner ihre Daten kontrollieren sowie deren Verwendung. Andererseits werden Daten so über Unternehmensgrenzen hinweg eher verfügbar gemacht, sodass insbesondere KMUs und Startups Zugang zu ihnen erhalten bei gleichzeitiger Übernahme von Verantwortung und Verfolgung des Ziels fairer KI. Faire KI bedeutet auch, dass Endkonsument*innen souverän mit ihr interagieren. Und das beginnt bereits bei der Art und Weise, wie Daten für das Training akquiriert werden. Die vorgestellten Mechanismen und Governance-Aspekte sollen einen möglichen Weg aufzeigen, wie die Souveränität von Endkonsument*innen gestärkt werden kann bei gleichzeitiger Stärkung von KMUs und Startups in ihrer Möglichkeit, Daten zu akquirieren und KI-basierte Systeme zu trainieren.

Acquisti, A. / Adjerid, I. / Balebako, R. / Brandimarte, L. / Cranor, L. F. / Komanduri, S., Leon, P. G. / Sadeh, N. / Schaub, F. / Sleeper, M. / Wang, Y. / Wilson, S. (2018): Nudges for Privacy and Security: Understanding and Assisting Users' Choices Online, in: ACM Computing Surveys, Jg. 50 / Nr. 3, 1–41, DOI: 10.1145/3054926.

Aitken, M. / de St. Jorre, J. / Pagliari, C. / Jepson, R. / Cunningham-Burley, S. (2016): Public Responses to the Sharing and Linkage of Health Data for Research Purposes: A Systematic Review and Thematic Synthesis of Qualitative Studies, in: BMC Medical Ethics, Jg. 17 / Nr. 1, 73, DOI: 10.1186/s12910-016-0153-x.

Brîncoveanu, C. / Carl, K. V. / Binz, S. / Weiher, M.-A. / Thomas, O. / Hinz, O. (2024): Artificial Intelligence-Based Assistance Systems for Environmental Sustainability in Smart Homes: A Systematic Literature Review on Requirements and Future Directions, in: Klein, M. / Krupka, D. / Winter, C. / Gergeleit, M. / Martin, L. (Hrsg.): INFORMATIK 2024. Lecture Notes in Informatics (LNI), Bonn: Gesellschaft für Informatik, 1165–1182.

Carl, K. V. / Hinz, O. (2024): What We Already Know About Corporate Digital Responsibility in IS Research: A Review and Conceptualization of Potential CDR Activities, in: Electronic Markets, Jg. 34 / Nr. 27, 1–30 DOI: 10.1007/s12525-024-00708-0.

COMET (2025): COMET – SmartLivingNEXT, URL: https://smartlivingnext.de/smartlivingnext-comet/ (aufgerufen am: 16/06/2025).

Dreesbach, T. / Dankert, L. / Ruhe, J. / Carl, K. V. / Hinz, O. / Thomas, O. (2025): Data Unlocked — Ein Modell zur anwendungsübergreifenden Nutzung persönlicher Daten im Smart-Living-Kontext, in: HMD Praxis der Wirtschaftsinformatik [im Erscheinen].

Feuerriegel, S. / Dolata, M. / Schwabe, G. (2020): Fair AI, in: Business & Information Systems Engineering, Jg. 62 / Nr. 4, 379–384, DOI: 10.1007/s12599-020-00650-3.

Hanneke, B. / Baum, L. / Schlereth, C. / Hinz, O. (2023): Consumer Preferences for Privacy Management Systems, in: De', R. / Paul, S. / Sarker, S. / Tuunainen, V. K. / Fernández W. D. / Nandhakumar, J. / Santhanam, R. (Hrsg.): Proceedings of the 44th International Conference on Information Systems, ICIS 2023, Rising Like a Phoenix: Emerging from the Pandemic and Reshaping Human Endeavors with Digital Technologies, Atlanta: Association for Information Systems.

Mihale-Wilson, C. / Zibuschka, J. / Carl, K. V. / Hinz, O. (2021): Corporate Digital Responsibility – Extended Conceptualization and a Guide to Implementation, in: ECIS 2021 Research Papers, Nr. 80.

Oliveira, M. I. S. / Barros Lima, G. de F. / Farias Lóscio, B. (2019): Investigations into Data Ecosystems: A Systematic Mapping Study, in: Knowledge and Information Systems, Jg. 61 / Nr. 2, 589–630, DOI: 10.1007/s10115-018-1323-6.

Scheider, S. / Lauf, F. / Möller, F. / Otto, B. (2023): A Reference System Architecture with Data Sovereignty for Human-Centric Data Ecosystems, in: Business & Information Systems Engineering, Jg. 65 / Nr. 5, 577–595, DOI: 10.1007/s12599-023-00816-9.

SmartLivingNext (2025): COMET – Consumer Empowerment und Teilhabe in der Domäne Smart Living, URL: https://smartlivingnext.de/smartlivingnext-comet/ (aufgerufen am: 10/07/2025).

Tallon, P. P. (2013).: Corporate Governance of Big Data: Perspectives on Value, Risk, and Cost, in: Computer, Jg. 46 / Nr. 6, 32–38, DOI: 10.1109/MC.2013.155.

van Houten, H. (2020): For Fair and Equal Healthcare, We Need Fair and Bias-Free AI, URL: https://www.philips.com/a-w/about/news/archive/blogs/innovation-matters/2020/20201116-for-fair-and-equal-healthcare-we-need-fair-and-bias-free-ai.html (aufgerufen am: 17/06/2025).

von Zahn, M. / Feuerriegel, S. / Kuehl, N. (2022): The Cost of Fairness in AI: Evidence from E-Commerce, in: Business & Information Systems Engineering, Jg. 64 / Nr. 3, 335–348, DOI: 10.1007/s12599-021-00716-w.

Witzki, A. / Carl, K. V. / Dankert, L. / Thomas, O. / Hinz, O. (2025a): What Are Your Intentions With My Data? A Tool to Enhance Consumer Data Sovereignty in Digital Ecosystems, in: Roßnagel, H. / Schunck, C. H. / Pöhn, D. (Hrsg.): OID 2025, Bonn: Gesellschaft für Informatik e.V., DOI: 10.18420/OID2025_01.

Witzki, A. / Carl, K. V. / Hinz, O. (2025b): Strategische Entscheidungsprozesse in digitalen Startups: Strukturierung von Faktoren für den Beitritt zu einem digitalen Ökosystem, in: HMD Praxis der Wirtschaftsinformatik; Jg. 62 / Nr. 1, 531–549, DOI: 10.1365/s40702-025-01155-x.

3.

Governance

Fairness als Wertschöpfungsfaktor[1]

Führung zwischen Innovation und Ethik

Maximilian Förster

1. Die doppelte Herausforderung von Innovation und Fairness

In einer Zeit tiefgreifender digitaler Transformation stehen Unternehmen vor der doppelten Herausforderung, einerseits innovative Technologien wie Künstliche Intelligenz (KI) zu nutzen, um wettbewerbsfähig zu bleiben, und andererseits ethische Grundsätze zu wahren, die das Vertrauen ihrer Stakeholder – und damit ihre eigene „license to operate" – sichern. Diese Spannung manifestiert sich im Kontext von Corporate Digital Responsibility (CDR) besonders in dem Wert „Fairness", der zunehmend als kritischer Erfolgsfaktor für nachhaltige Wertschöpfung erkannt wird. Die rasante Entwicklung von KI-Technologien hat tiefgreifende Auswirkungen auf Geschäftsmodelle, Arbeitsprozesse und Kundenbeziehungen. Algorithmen treffen heute Entscheidungen, die früher Menschen vorbehalten waren – von der Personalauswahl über die Kreditvergabe bis zur Preisgestaltung. Diese Automatisierung verspricht Effizienzgewinne, birgt jedoch das Risiko, bestehende Ungleichheiten zu verstärken oder neue Formen der Diskriminierung zu schaffen. Die Sicherstellung von Fairness in algorithmischen Entscheidungssystemen stellt deshalb sowohl eine ethische als auch wirtschaftliche Notwendigkeit dar. Hier setzt Corporate Digital Responsibility an, ein Konzept, das den verantwortungsvollen Umgang mit digitalen Technologien und Daten in Unternehmen adressiert. CDR ist, wie Alexander Brink hervorhebt, „nicht nur ein ethischer Imperativ, sondern auch eine kluge unternehmerische Entscheidung" (CDR-Initiative 2025). Die wirtschaftsethische Dimension von CDR rückt dabei zunehmend in den Fokus: Fairness wird nicht mehr nur als moralische

[1] Generative KI wurde ausschließlich für sprachliche Überarbeitungen und Formulierungsvorschläge eingesetzt. Sämtliche inhaltlichen Aussagen, Analysen und Schlussfolgerungen stammen vom Autor und wurden redaktionell geprüft.

Verpflichtung betrachtet, sondern als strategischer Wertschöpfungsfaktor, der Vertrauen schafft, Reputation stärkt und langfristige Kundenbeziehungen fördert (vgl. Suchanek 2020: 18 f.).

Trotz dieser Erkenntnis mangelt es vielen Unternehmen an konkreten Ansätzen, Fairness systematisch in ihre Managementprozesse zu integrieren. Die naheliegende Frage lautet daher: Wie können Unternehmen Fairness im Zeitalter der Künstlichen Intelligenz als Wertschöpfungsfaktor in ihre Management-Modelle integrieren und dabei die Balance zwischen Innovation und ethischer Verantwortung wahren? Um diese zu beantworten, werden zwei etablierte Management-Modelle analysiert: das St. Galler Management-Modell mit seiner systemischen Perspektive und das Leipziger Führungsmodell mit seinem Fokus auf werteorientierte Führung. Anschließend wird ein holistischer, beide Perspektiven verknüpfender Ansatz entwickelt. Anhand von Fallbeispielen aus der Unternehmenspraxis – darunter die Allianz Versicherung und Merck – werden konkrete Implementierungsstrategien aufgezeigt und Implikationen für die Managementpraxis abgeleitet.

2. Das St. Galler Management Modell: Systemische Perspektive auf Fairness

2.1 Grundlagen und systemische Sichtweise des St. Galler Management Modells

Das St. Galler Management Modell (SGMM) hat sich seit seiner ersten Konzeption in den 1960er Jahren zu einem einflussreichen integrativen Management-Ansatz entwickelt (vgl. Kämpf-Dern 2022: 36). In seiner aktuellen Version, entwickelt von Johannes Rüegg-Stürm und Simon Grand, bietet es eine systemorientierte Perspektive auf Unternehmensführung, die besonders geeignet ist, um komplexe Herausforderungen wie die Integration von Fairness-Prinzipien in digitale Transformationsprozesse zu adressieren. Das SGMM versteht Organisationen als komplexe, in ihre Umwelt eingebettete Systeme, deren Hauptaufgabe in der Schaffung von Wert liegt. Diese systemische Sichtweise ermöglicht es, Fairness nicht isoliert, sondern in ihren vielfältigen Wechselwirkungen mit anderen Aspekten der Unternehmensführung zu betrachten. Das Management in unserer heutigen komplexen Welt ist darauf angewiesen, Interdependenzen zu erkennen und zu gestalten – ein Ansatz, der für die Verankerung von Fairness im Kontext von KI-Anwendungen besonders wertvoll ist (vgl. Rüegg-Stürm/Grand 2020: 37 f.). Zentral für das SGMM ist die Unterscheidung zwischen Aufgaben- und Praxisperspektive. Die Aufgabenperspektive, beschreibt sechs Schlüsselkategorien: Umweltsphären, Stakeholder, Interaktionsthemen, Ordnungsmomente, Prozesse

und Entwicklungsmodi. Die Praxisperspektive hingegen fokussiert auf die konkrete Managementpraxis und die Gestaltung von Kommunikation. Beide Perspektiven sind für die Integration von Fairness-Prinzipien relevant, da sie strukturelle und kulturelle Aspekte umfassen (vgl. ebd.: 41–44).

2.2 Die Aufgabenperspektive des SGMM: Strukturelle und prozessuale Verankerung von Fairness

Die Aufgabenperspektive des SGMM – dargestellt in Abbildung 1 – verortet Fairness an den Schnittstellen technologischer, wirtschaftlicher und gesellschaftlicher Umweltsphären und betont die Notwendigkeit, die oft widersprüchlichen Erwartungen verschiedener Anspruchsgruppen an faire KI-Systeme zu verstehen und in Einklang zu bringen (vgl. ebd.: 56–65). Wie das Modell hervorhebt, ist die „Umwelt nicht als gegeben vorausgesetzt […], sondern sie wird von einer Organisation [in kommunikativen Prozessen] spezifisch verfertigt" (ebd.: 237) – eine wichtige Erkenntnis für die partizipative Gestaltung fairer KI-Systeme.

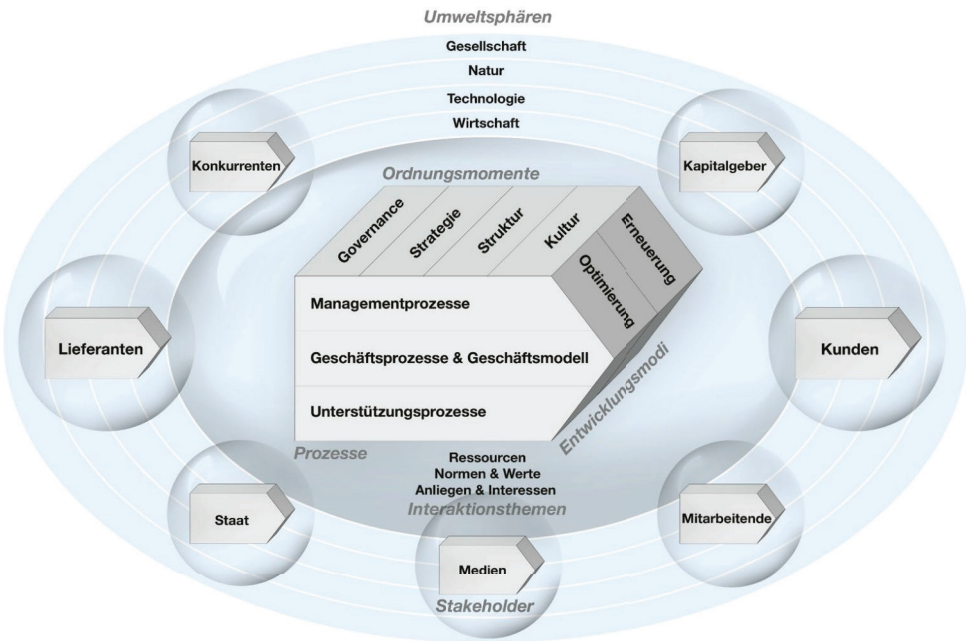


ABBILDUNG 1: DIE AUFGABENPERSPEKTIVE DES ST. GALLER MANAGEMENT-MODELLS (QUELLE: RÜEGG-STÜRM/GRAND 2020: 53)

Die Ordnungsmomente – Governance, Strategie, Strukturen und Kultur – bieten konkrete Ansatzpunkte: von ethischen Leitplanken und Verantwortlichkeiten (Governance) über die Positionierung von Fairness als strategisches Asset (Strategie) und die Anpassung organisationaler Strukturen (Struktur) bis zur Kultivierung einer von Fairness geprägten Unternehmenskultur (Kultur) (vgl. ebd.: 87–109). Fairness kann als strategischer Wettbewerbsvorteil und Differenzierungsmerkmal dienen. Stakeholder*innen, insbesondere Kund*innen und Mitarbeitende, erwarten zunehmend faire KI-Systeme (vgl. Teodorescu/Makridis 2024). Eine Kultur, die Werte (wie z. B. Fairness) bewusst ausprägt, fördert intrinsisch motiviertes ethisches Verhalten und die Entwicklung sowie den Einsatz fairer KI-Systeme, wodurch der Bedarf an rigider Kontrolle reduziert werden kann (vgl. Brink 2022: 8). Unternehmen, die diesen Erwartungen gerecht werden, können Vertrauen aufbauen, was wiederum eine höhere Kund*innentreue und besserer Mitarbeitendenbindung verspricht.

Eine rein technische oder nur auf Governance fokussierte Implementierung greift also zu kurz. Ohne eine tiefgreifende kulturelle Verankerung und entsprechende Anpassungen in Strategie und Struktur bleiben Fairness-Bemühungen oberflächlich. Es bedarf eines holistischen Ansatzes, der alle Ordnungsmomente berücksichtigt, um eine systemische und sich gegenseitig verstärkende Wirkung zu erzielen.

Auch die Prozesse einer Organisation müssen unter CDR-Gesichtspunkten gestaltet werden, wobei Prinzipien wie Ethics by Design und Privacy by Design bereits früh in der Produktentwicklung integriert werden sollten (vgl. Spiekermann 2019: 276–279). Dieser präventive Ansatz stellt einen Paradigmenwechsel von reaktiver Problembehebung zu proaktiver Werteintegration dar. Die Entwicklungsmodi des SGMM, Optimierung (kontinuierliche Verbesserung bestehender KI-Systeme) und Erneuerung (Neuentwicklung inhärent fairer Lösungen), bieten dabei Ansätze zur fairen Gestaltung digitaler Wertschöpfungsprozesse, was Unternehmen vor strategische Entscheidungen zwischen inkrementeller Anpassung und fundamentalem Neuentwurf stellt (vgl. Rüegg-Stürm/Grand 2020: 110–117).

2.2 Die Praxisperspektive – kulturelle und kommunikative Voraussetzungen für Fairness

Die Praxisperspektive des SGMM, welche in Abbildung 2 dargestellt ist, erweitert die Aufgabenperspektive, indem sie soziokulturelle und kommunikative Voraussetzungen für eine wirksame Wertschöpfung und damit auch für gelebte Fairness fokussiert, die über formale Regeln hinausgehen. Management wird als „reflexive Gestaltungspraxis" verstanden, was bedeutet, dass Fairness in KI-Systemen eine lebendige Kommunikationskultur erfordert, in der ethische Fragen offen diskutiert werden (vgl. ebd.: 122 f.).

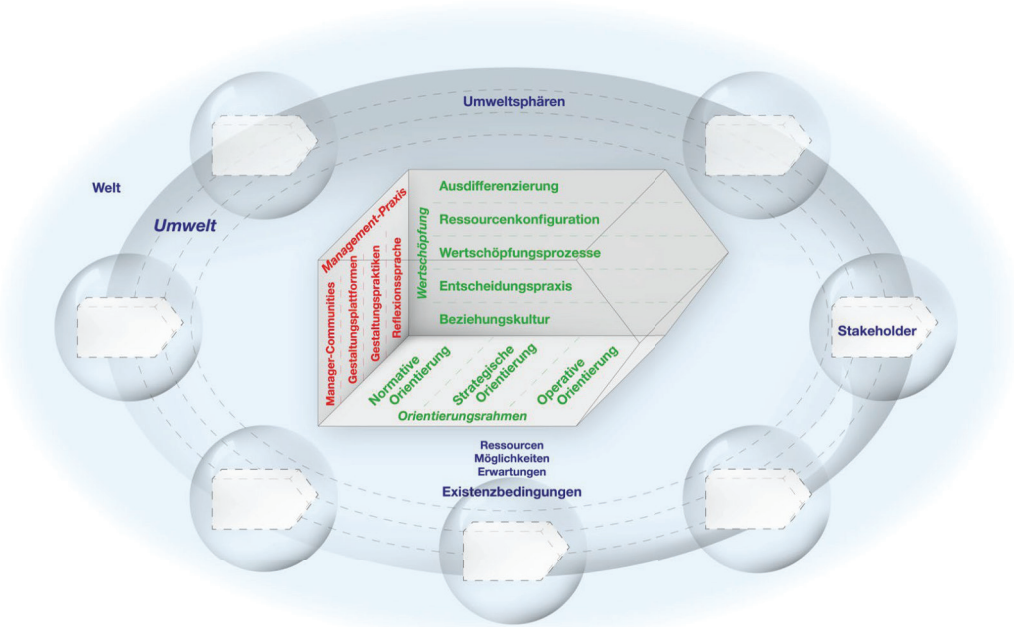


ABBILDUNG 2: DIE PRAXISPERSPEKTIVE DES ST. GALLER MANAGEMENT-MODELLS (QUELLE: RÜEGG-STÜRM/GRAND 2020: 159)

Wertschöpfung wird als ko-kreativer Prozess verstanden, bei dem faire KI nicht allein durch technisch einwandfreie Algorithmen, sondern vor allem durch Gestaltungsprozesse entsteht, die faire und vielfältige Perspektiven einbeziehen (vgl. ebd.: 166–174). Die Art und Weise, wie Fairness im Unternehmen verstanden und umgesetzt wird, ist somit weniger ein feststehendes technisches Attribut als vielmehr ein Ergebnis kontinuierlicher sozialer Aushandlungsprozesse. Damit Fairness wirksam wird, muss sie Teil eines geteilten Orientierungsrahmens sein und in täglichen Routinen verankert werden, was kontinuierliche Kommunikation und Reflexion erfordert (vgl. ebd.: 206–220). Die Umwelt wird dabei als ein durch Interaktionen und Deutungen mitgestalteter Möglichkeitsraum begriffen, in dem gesellschaftliche Erwartungen an faire KI aktiv berücksichtigt werden (vgl. ebd.: 222–237). Führungskräfte und Entwicklungsteams müssen Räume für kritische Reflexion schaffen, um Annahmen zu hinterfragen und ethische Konflikte frühzeitig zu erkennen – ein Kernmechanismus, um unbewusste Vorurteile in KI-Systemen zu identifizieren und zu mitigieren (vgl. ebd.: 238–272). Die Praxisperspektive verdeutlicht, dass die Verankerung von Fairness über Regeln und Strukturen hinausgehen muss und eine Kultur des Hinsehens, Dialogs und gemeinsamer Verantwortungsübernahme erfordert, in der Fairness zu einer selbstverständlichen Komponente des täglichen Handelns und Entscheidens wird.

2.3 Das SGMM im Kontext werteorientierter Führung

Das St. Galler Management Modell bietet mit seiner systemischen Perspektive einen wertvollen Rahmen für die Verankerung von Fairness im Kontext der digitalen Transformation. Es ermöglicht ein ganzheitliches Verständnis der Wechselwirkungen zwischen technologischen, wirtschaftlichen und gesellschaftlichen Aspekten und bietet konkrete Ansatzpunkte für die strategische, strukturelle und kulturelle Integration von Fairness-Prinzipien. Die besondere Stärke des Modells liegt in seiner Fähigkeit, Komplexität zu erfassen und zu gestalten – eine Eigenschaft, die angesichts der vielschichtigen Herausforderungen fairer KI-Systeme von unschätzbarem Wert ist. Gleichzeitig sind die Prinzipien einer werteorientierten Führung, die nach einem Purpose fragt, sowie den Unternehmergeist ins Zentrum stellt, in diesem Modell tendenziell unterbeleuchtet. Das Leipziger Führungsmodell kann diese Lücke schließen und zu einer fairen Unternehmensführung im Zeitalter von Künstlicher Intelligenz beitragen.

3. Das Leipziger Führungsmodell: Werteorientierte Führung und Fairness

3.1 Grundlagen und die vier Dimensionen des Leipziger Führungsmodells

Das Leipziger Führungsmodell (LFM), entwickelt an der HHL Leipzig Graduate School of Management, bietet eine komplementäre Perspektive zum St. Galler Management Modell, indem es den Fokus auf werteorientierte Führung legt. Es adressiert die Frage, was „gute Führung" im 21. Jahrhundert ausmacht, und ist damit besonders relevant für die Verankerung von Fairness-Prinzipien in Führungsentscheidungen im Kontext Künstlicher Intelligenz. Im Kern des LFMs stehen vier miteinander verbundene Dimensionen: Purpose (Sinn und Zweck), Unternehmergeist, Verantwortung und Effektivität. Diese Dimensionen bilden ein Spannungsfeld, in dem sich Führungskräfte bewegen und Entscheidungen treffen müssen. Für die Integration von Fairness in KI-gestützte Geschäftsprozesse bietet dieses mehrdimensionale Verständnis von Führung wertvolle Ansatzpunkte (vgl. Kirchgeorg et al. 2019: 15–16). Das Leipziger Führungsmodell ist in Abbildung 3 dargestellt.

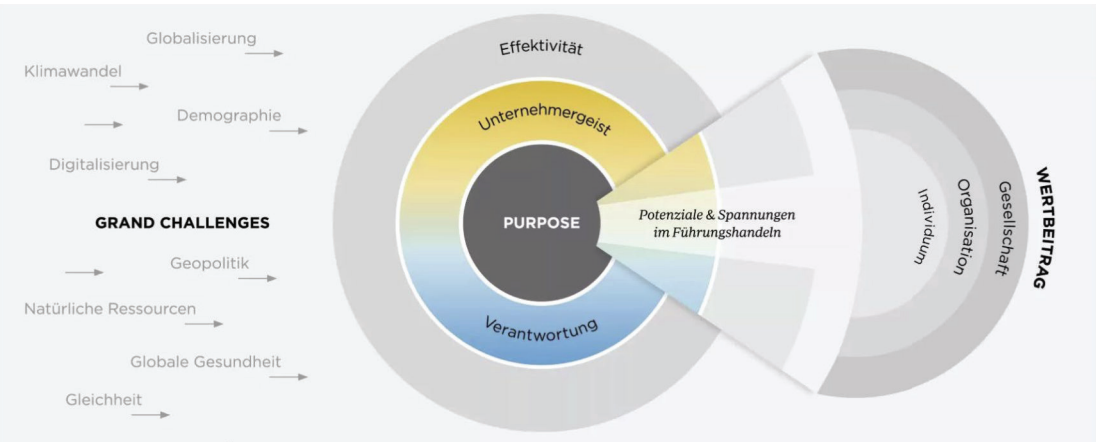


ABBILDUNG 3: DAS LEIPZIGER FÜHRUNGSMODELL
(QUELLE: KIRCHGEORG ET AL. 2019: 15)

3.2 Purpose und Public Value im Kontext von Fairness

Die Purpose-Dimension des Modells fragt nach dem „Warum" unternehmerischen Handelns und betont, dass erfolgreiche Führung einen Beitrag zu einem größeren Ganzen leisten muss.

> Gute Führung bedeutet im Leipziger Führungsmodell, einen Beitrag zu einem größeren Ganzen zu leisten, den Dritte als sinn- und wertvoll erachten, ohne dass legitime Rechte und Ansprüche Dritter verletzt werden (ebd.: 5).

Diese Definition unterstreicht zentrale Fairnessaspekte durch die Forderung, Rechte und Ansprüche aller Betroffenen zu berücksichtigen. Das Konzept des Public Value, ein Kernbestandteil des Leipziger Führungsmodells, ist hierbei besonders relevant. Es definiert den gesellschaftlichen Nutzen und Wertbeitrag einer Organisation zum Gemeinwohl. Auf KI-Systeme übertragen bedeutet dies, Fairness als gesamtgesellschaftlichen Wertbeitrag statt lediglich als internes Prinzip zu begreifen. Das Modell betont, dass die „license to operate" von Unternehmen immer stärker von ihrem glaubwürdigen Beitrag zum Gemeinwohl abhängt – ein wichtiger Aspekt für die Legitimation algorithmischer Entscheidungssysteme (vgl. ebd.: 17–23). Fairness wird in diesem Zusammenhang zu einem integralen Bestandteil der Unternehmensführung, da diese die gesellschaftliche Akzeptanz und damit die Handlungsfähigkeit des Unternehmens sichert.

3.3 Unternehmergeist und Innovation für faire KI-Systeme

Die Dimension des Unternehmergeistes im Leipziger Führungsmodell betont die Notwendigkeit von Innovation, Kreativität und Risikobereitschaft. Im Kontext der KI-Entwicklung steht diese Dimension in einem produktiven Spannungsverhältnis zur Fairness: Einerseits erfordert die Entwicklung fairer Algorithmen innovative Ansätze und unternehmerisches Denken, andererseits müssen Innovationen stets auf ihre Fairness-Implikationen hin geprüft werden. Das Modell spricht hier von der „ständigen Balance zwischen dem Hervorbringen von Inventionen (Exploration) und der möglichst effizienten und effektiven Umsetzung bewährter Produkte und Prozesse (Exploitation)" (ebd.: 31) – eine Balance, die auch für die Entwicklung fairer KI-Systeme entscheidend ist (vgl. ebd.: 24–33).

Die Effektivitätsdimension des Modells ergänzt diese Perspektive, indem sie betont, dass es nicht nur um Effizienz (das richtige Tun), sondern vor allem um Effektivität (das Richtige tun)

geht. Für die Verankerung von Fairness bedeutet dies, dass es nicht ausreicht, Fairness-Prinzipien zu formulieren – sie müssen auch wirksam in die Praxis umgesetzt werden. Das Modell betont, dass „effizient ausgerichtete Strategien in sich wandelnden Kontexten oft nicht mehr das richtige Ziel ansteuern" (ebd.: 41), was angesichts der dynamischen Entwicklung von KI und gesellschaftlichen Fairness-Erwartungen besonders bedeutsam ist (vgl. ebd.: 41–47).

3.4 Verantwortung, ethische Führung und Stimmigkeit

Die Verantwortungsdimension des Leipziger Führungsmodells adressiert direkt ethische Fragen der Unternehmensführung. Verantwortung wird dabei nicht nur als Haftung für negative Folgen verstanden, sondern als proaktive Gestaltungsaufgabe. Für die Integration von Fairness in KI-Systeme bedeutet dies, dass Führungskräfte aktiv Verantwortung für die Gestaltung fairer Algorithmen übernehmen müssen, anstatt nur auf rechtliche Anforderungen oder öffentlichen Druck zu reagieren. Das Modell betont zudem die Vorbildfunktion von Führungskräften – ein Aspekt, der für die Etablierung einer Fairness-Kultur im Unternehmen von entscheidender Bedeutung ist. Ein zentrales Merkmal des Leipziger Führungsmodells ist sein Fokus auf Stimmigkeit (Konsistenz). Es geht darum, die vier Dimensionen zusammenzudenken und ein kohärentes Führungsverständnis zu entwickeln (vgl. ebd.: 34–40).

Für die Integration von Fairness ist es demnach erforderlich, Fairness-Prinzipien nicht isoliert zu betrachten, sondern dass dies im Einklang mit dem Purpose des Unternehmens, seinem unternehmerischen Handeln, seiner Verantwortung und seiner Effektivität stehen müssen. Diese ganzheitliche Perspektive hilft, Fairness nicht als „Add-on", sondern als integralen Bestandteil der Unternehmensführung zu verstehen.

Zudem betont das Leipziger Führungsmodell die Bedeutung eines „inneren Kompasses" für Führungskräfte. „Wer sich selbst nicht führen kann, kann auch andere nicht führen" (ebd.: 11) stellt eine der Grundannahmen dar. Für die Verankerung von Fairness in KI-Systemen bedeutet dies, dass Führungskräfte ein tiefes Verständnis und eine persönliche Überzeugung von der Bedeutung von Fairness entwickeln müssen, um diese Werte glaubwürdig im Unternehmen zu vertreten und zu implementieren.

Das Leipziger Führungsmodell mit seinem Fokus auf werteorientierte Führung stellt eine wertvolle Ergänzung zur systemischen Perspektive des St. Galler Management Modells dar. Es

adressiert die normativen Aspekte der Unternehmensführung und bietet konkrete Ansatzpunkte für die Verankerung von Fairness in Führungsentscheidungen im Kontext der Künstlichen Intelligenz. Seine Stärke liegt in der Verbindung ethischer Prinzipien mit unternehmerischem Handeln, was einen Rahmen für verantwortungsvolle Innovation schafft.

4. Integration der Modelle: Ein holistischer Ansatz

4.1 Komplementarität der Modelle und CDR als Integrationsrahmen

Die Analyse des St. Galler Management Modells und des Leipziger Führungsmodells zeigt, dass beide Ansätze wertvolle, aber unterschiedliche Perspektiven auf die Verankerung von Fairness im Unternehmenskontext bieten. Während das St. Galler Modell mit seiner systemischen Sicht die strukturellen und prozessualen Aspekte unterstreicht, fokussiert das Leipziger Modell auf die werteorientierte Führung und die normative Dimension. Ein holistischer Ansatz zur Integration von Fairness als Wertschöpfungsfaktor im Zeitalter der Künstlichen Intelligenz sollte beide Perspektiven verbinden. Die Komplementarität der beiden Modelle zeigt sich besonders deutlich in ihren unterschiedlichen Betrachtungsebenen. Das St. Galler Management Modell nimmt eine Makroperspektive ein, indem es Organisationen als komplexe, in ihre Umwelt eingebettete Systeme betrachtet und die Wechselwirkungen zwischen verschiedenen Umweltsphären, Anspruchsgruppen und organisationalen Elementen analysiert. Das Leipziger Führungsmodell hingegen bietet eine Mikroperspektive, die auf die Rolle und Verantwortung individueller Führungskräfte fokussiert und konkrete Handlungsorientierungen für wertebasierte Entscheidungen gibt.

Corporate Digital Responsibility (CDR) bietet einen geeigneten Integrationsrahmen für diese beiden Perspektiven. CDR bezieht sich nach Dörr (2025: 4):

> [E]inerseits auf die Beachtung digitaler Nachhaltigkeit und anderseits auf Berücksichtigung der sozialen, ökonomischen und ökologischen Wirkungen digitalen Unternehmenshandelns in der Welt. Dabei handelt es sich um freiwillige unternehmerische Aktivitäten, die über das gesetzlich Vorgeschriebene hinausgehen. Sie kann für Fairness sorgen und die digitale Transformation zum gemeinsamen Vorteil aller sowie einer nachhaltigen Entwicklung gestalten helfen.

Diese Definition verdeutlicht, dass CDR sowohl systemische Aspekte (soziale, ökonomische und ökologische Auswirkungen) als auch Führungsaspekte (freiwillige unternehmerische Aktivitäten) umfasst. Zudem wird explizit der Fairness-Aspekt genannt.

Die wirtschaftsethische Dimension von CDR lässt sich durch die Integration der beiden Modelle besonders gut adressieren. Wirtschaftsethik fragt nach der moralischen Qualität wirtschaftlichen Handelns und sucht nach Wegen, wie ökonomische Effizienz und ethische Verantwortung in Einklang gebracht werden können (vgl. Ulrich 2008: 101–102). Genau dieses Spannungsfeld wird durch die Kombination des systemorientierten St. Galler Modells mit dem werteorientierten Leipziger Führungsmodell erfasst. Fairness als Kernprinzip von CDR nimmt somit im Kontext algorithmischer Entscheidungssysteme eine entscheidende Rolle ein.

4.2 Ein integriertes Fairness-Management-Modell

Ein integriertes Fairness-Management-Modell sollte drei Ebenen der Verankerung berücksichtigen, um Risiken bei der KI-Implementierung zu minimieren: die strukturelle, die kulturelle und die prozessuale Ebene vgl. Nuhić-Mešković/Mešković 2023: 13 .

Auf struktureller Ebene geht es um die Schaffung organisationaler Rahmenbedingungen für faire KI-Systeme – von der Etablierung spezialisierter Rollen und Verantwortlichkeiten bis hin zur Integration von Fairness-Kriterien in Governance-Strukturen. Das St. Galler Modell bietet hier mit seinen Ordnungsmomenten (Strategie, Strukturen, Kultur) wertvolle Orientierung. Auf kultureller Ebene steht die Verankerung von Fairness-Werten in der Unternehmenskultur im Mittelpunkt. Hier bietet das Leipziger Führungsmodell mit seinem Fokus auf Purpose und Verantwortung wichtige Ansatzpunkte. Wie Spiekermann (vgl. 2019: 273) ausführt, reicht es nicht aus, digitalethische Prinzipien zu formulieren – sie müssen gelebt und in der täglichen Praxis umgesetzt werden. Dies erfordert Führungskräfte mit einem „inneren Kompass“, die Fairness authentisch vorleben und fördern. Die prozessuale Ebene betrifft die Integration von Fairness in konkrete Arbeitsabläufe und Entscheidungsprozesse. Hier bietet das St. Galler Modell mit seiner Prozessorientierung wertvolle Ansätze, während das Leipziger Modell mit seinem Fokus auf Effektivität die zielgerichtete Umsetzung unterstützt. Konkret kann dies bedeuten, Fairness-Kriterien in den gesamten Lebenszyklus von KI-Systemen zu integrieren – von der Problemdefinition über die Datenauswahl und Modellentwicklung bis hin zur Implementierung und kontinuierlichen Überwachung.

4.3 Herausforderungen und Lösungsansätze bei der Implementierung

Die Integration der beiden Modelle stößt jedoch auch auf Herausforderungen. Eine zentrale Herausforderung sind Zielkonflikte zwischen Innovation und Fairness. Während das St. Galler Modell mit seinem Fokus auf Wertschöpfung tendenziell die Innovationsperspektive betont, hebt das Leipziger Modell mit seinem Fokus auf Verantwortung eher die ethische Dimension hervor. Ein integrierter Ansatz muss diese Spannung produktiv nutzen, indem er Fairness nicht als Hindernis, sondern als Treiber für verantwortungsvolle Innovation versteht. Eine weitere Herausforderung sind organisationale Widerstände gegen die Verankerung von Fairness-Prinzipien, etwa aus Kosten- oder Komplexitätsbedenken oder unternehmenskulturellen Gründen. Hier bietet die Kombination der beiden Modelle wertvolle Ansätze: Das St. Galler Modell hilft, systemische Widerstände zu identifizieren und zu adressieren, während das Leipziger Modell Orientierung für den Umgang mit individuellen Widerständen gibt. Schließlich stellen sich technische und methodische Herausforderungen bei der Operationalisierung von Fairness in KI-Systemen. Die verschiedenen, teils widersprüchlichen Fairness-Definitionen (vgl. Saxena et al. 2020: 2–3) erfordern einen integrierten Ansatz, der diese Komplexität anerkennt und flexible, kontextspezifische Lösungen ermöglicht, die systemischen Anforderungen (SGMM) und ethischen Prinzipien (LFM) gerecht werden.

Lösungsansätze für diese Herausforderungen umfassen die Entwicklung von Fairness-Frameworks, die verschiedene Fairness-Definitionen berücksichtigen, die Etablierung cross-funktionaler Teams, die technische und ethische Expertise verbinden, sowie die kontinuierliche Evaluation und Anpassung von Fairness-Maßnahmen im Sinne eines iterativen Lernprozesses.

4.4 Zukunftsperspektiven für Fairness als Wertschöpfungsfaktor

Fairness etabliert sich immer mehr als strategischer Erfolgsfaktor. Angesichts der wachsenden Dominanz von KI-Systemen bei Entscheidungsprozessen wird die erfolgreiche Integration von Fairness zum entscheidenden Wettbewerbsvorteil. Unternehmen, die Fairness nicht nur als ethische Verpflichtung, sondern als Wertschöpfungsfaktor verstehen und in ihre Geschäftsmodelle integrieren, werden besser in der Lage sein, das Vertrauen ihrer Stakeholder zu gewinnen und langfristig erfolgreich zu sein.

Die Einbettung von Fairness-Prinzipien in bestehende Management-Frameworks wird dabei eine zentrale Rolle spielen. Trotz der bestehenden Herausforderungen bieten beide Management-Modelle vielversprechende Ansätze für die Verstetigung von Fairness als Wertschöpfungsfaktor im Zeitalter der Künstlichen Intelligenz. Durch die Verbindung der systemischen Perspektive des St. Galler Modells mit der werteorientierten Perspektive des Leipziger Modells entsteht ein holistischer Rahmen, der sowohl die strukturellen als auch die kulturellen Voraussetzungen für faire KI-Systeme adressiert und so einen Beitrag zur verantwortungsvollen Gestaltung der digitalen Transformation leistet.

Die theoretischen Erkenntnisse aus der Integration des St. Galler Management Modells und des Leipziger Führungsmodells lassen sich anhand konkreter Unternehmensbeispiele veranschaulichen. Im Folgenden werden zwei Fallbeispiele vorgestellt, die unterschiedliche Aspekte der Verankerung von Fairness als Wertschöpfungsfaktor im Kontext der Künstlichen Intelligenz illustrieren: die Allianz Versicherung und Merck. Aus diesen Beispielen werden anschließend Implikationen für die Managementpraxis abgeleitet.

5. Fallbeispiele und Implikationen

5.1 Allianz: Fairness in der KI-gestützten Versicherungsbranche

Die Allianz Versicherung hat sich in den letzten Jahren intensiv mit der digitalen Transformation ihres Geschäftsmodells beschäftigt und dabei Fairness als zentralen Wert verankert. Wie Lucie Bakker, Vorständin für Schaden bei der Allianz, betont, verändert sich die Versicherungsbranche grundlegend durch den Einsatz von KI: „Wir haben uns vorgenommen, deutlich schneller zu werden“ (Bakker, zitiert nach Welp 2024). Gleichzeitig ist die Versicherungsbranche besonders anfällig für Fairness-Probleme, da algorithmische Entscheidungen direkte finanzielle Auswirkungen auf Kunden haben können (vgl. Kovacevic 2024).

Die Allianz hat einen systemischen Ansatz gewählt, der Elemente des St. Galler Management Modells widerspiegelt. Sie hat KI zunächst dem abgegrenzten Bereich der Tier-Krankenversicherung implementiert und dort Erfahrungen gesammelt, bevor die Technologie auf andere Bereiche ausgeweitet wurde (vgl. Welp 2024). Dieses schrittweise Vorgehen entspricht dem im St. Galler

Modell aufgezeigten Zusammenspiel von Optimierung und Erneuerung und ermöglicht eine kontinuierliche Anpassung der Fairness-Kriterien. Gleichzeitig zeigt das Beispiel der Allianz auch Elemente des Leipziger Führungsmodells, insbesondere in Bezug auf die Balance zwischen Effektivität und Verantwortung. Laut Allianz soll der Einsatz von KI nicht zum Personalabbau führen, sondern zur Entlastung von Routineaufgaben (vgl. Kovacevic 2024). Diese Haltung spiegelt die im LFM benannte Verantwortung gegenüber Mitarbeitenden wider.

Besonders bemerkenswert ist der Umgang der Allianz mit dem Spannungsfeld zwischen Automatisierung und menschlicher Expertise. Während einfache Schadenfälle vollautomatisch bearbeitet werden können, betont das Unternehmen die Bedeutung menschlicher Entscheidungen bei komplexen Fällen: „Bei Überflutungen oder anderen größeren Schäden werden wir immer vor Ort sein und begutachten müssen" (Bakker, zitiert nach Welp 2024). Diese differenzierte Herangehensweise entspricht dem im St. Galler Modell betonten Prinzip der Selektivität und dem im Leipziger Modell hervorgehobenen Wertbeitrag.

5.2 Merck: Ethische Leitlinien für digitale Technologien

Merck hat mit seinem „Code of Digital Ethics" einen umfassenden Rahmen für den ethischen Umgang mit digitalen Technologien geschaffen, der auf fünf Kernprinzipien basiert: Autonomie, Nicht-Schädigung (Non-Maleficence), Wohltätigkeit (Beneficence), Gerechtigkeit (Justice) und Transparenz. (vgl. Becker et al. 2022: 2630–2635). Wie das Unternehmen betont, ist „digitale Ethik, also der ethische Umgang mit Daten und Algorithmen, von entscheidender Bedeutung. Unser Ziel? Die Chancen des digitalen Fortschritts nutzen und gleichzeitig die Risiken minimieren" (Merck KGaA 2025).

Mercks Ansatz spiegelt Elemente des St. Galler Management Modells wider, insbesondere in Bezug auf die Integration von Ethik in die Unternehmensstrategie. Das Unternehmen hat ein Digital Ethics Advisory Panel eingerichtet (DEAP), welches ethische Fragen im Zusammenhang mit digitalen Technologien bewertet und Empfehlungen ausspricht (vgl. Becker et al. 2022: 2630; Nemat et al. 2023: 742). Diese strukturelle Verankerung entspricht den im SGMM verankerten Ordnungsmomenten. Gleichzeitig zeigt Mercks Ansatz auch Elemente des Leipziger Führungsmodells, insbesondere in Bezug auf die Verantwortungsdimension. Das DEAP bewertet Risikominderungsvorschläge, dient als Forum für Richtlinien und berät proaktiv zu neuen Entwicklungen im Bereich digitaler

Ethik (vgl. ebd.: 742). Diese proaktive Herangehensweise entspricht der im LFM betonten Verantwortung als Gestaltungsaufgabe. Besonders bemerkenswert ist Mercks Principle-at-Risk Analysis (PaRA), ein Tool zur Operationalisierung digitaler Ethik. Dieses Tool dient dazu, „die Arbeit des Digital Ethics Advisory Panel bei dem multinationalen Wissenschafts- und Technologieunternehmen Merck KGaA zu leiten und zu harmonisieren" (ebd.: 741). Mercks Ansatz unterstreicht die Bedeutung des Ordnungsmoments „Governance" (Ethik-Kodex, Beratungsgremium) im SGMM und der Dimension „Verantwortung" im Leipziger Führungsmodell als zentrale Ankerpunkte für ethisches Handeln im digitalen Zeitalter.

5.3 Implikationen für die Managementpraxis

Aus den theoretischen Erkenntnissen und den Fallbeispielen lassen sich konkrete Handlungsempfehlungen für Führungskräfte ableiten, die Fairness als Wertschöpfungsfaktor in ihre Management-Modelle integrieren möchten.

Erstens sollten Führungskräfte Fairness nicht als isoliertes Prinzip, sondern als integralen Bestandteil ihrer Unternehmensstrategie betrachten. Wie das Beispiel von Merck zeigt, kann die Verankerung von Fairness in strategischen Dokumenten wie einem Code of Digital Ethics dazu beitragen, ethische Prinzipien in konkrete Handlungsanweisungen zu übersetzen. Diese strategische Verankerung entspricht sowohl dem St. Galler Modell (Strategie als Ordnungsmoment) als auch dem Leipziger Modell (Purpose als Orientierung).

Zweitens ist die Schaffung geeigneter organisationaler Strukturen entscheidend. Wie Merck zeigt, können spezialisierte Teams oder Gremien wie ein Digital Ethics Advisory Panel dazu beitragen, Fairness-Prinzipien in der Organisation zu verankern und kontinuierlich weiterzuentwickeln. Diese strukturelle Verankerung entspricht dem im St. Galler Modell betonten Zusammenspiel von Strategie, Strukturen und Kultur.

Drittens ist die Balance zwischen Automatisierung und menschlicher Expertise entscheidend. Wie das Beispiel der Allianz zeigt, sollte KI nicht als Ersatz, sondern als Ergänzung menschlicher Entscheidungen verstanden werden. Diese Balance entspricht dem im Leipziger Modell betonten Zusammenspiel von Unternehmergeist und Verantwortung und dem im St. Galler Modell hervorgehobenen Zusammenspiel von Optimierung und Erneuerung.

Viertens sollten Führungskräfte Fairness als kontinuierlichen Lernprozess verstehen. Wie die beiden Fallbeispiele zeigen, erfordert die Verankerung von Fairness in KI-Systemen ein kontinuierliches Monitoring, Evaluieren und Anpassen. Diese prozessuale Perspektive entspricht dem im St. Galler Modell ausgeprägten Entwicklungsmodus und dem im Leipziger Modell hervorgehobenen Lernaspekt von Führung.

Diese aus den theoretischen Erkenntnissen und den Fallbeispielen gewonnenen Handlungsempfehlungen verdeutlichen, wie Führungskräfte Fairness als Wertschöpfungsfaktor aktiv in ihre Management-Modelle integrieren können. Nach diesen praktischen Aspekten sollen im Folgenden die zentralen Erkenntnisse dieser Arbeit abschließend zusammengefasst und bewertet werden.

6. Fazit

Der vorliegende Beitrag hat aufgezeigt, wie Fairness als Wertschöpfungsfaktor in Management-Modelle integriert werden kann, um Unternehmen bei der Bewältigung der doppelten Herausforderung von Innovation und ethischer Verantwortung im Zeitalter der Künstlichen Intelligenz zu unterstützen. Durch die Analyse des St. Galler Management Modells und des Leipziger Führungsmodells sowie deren Integration im Rahmen der Corporate Digital Responsibility wurden konkrete Ansatzpunkte für die Verankerung von Fairness in Unternehmensstrukturen, -prozessen und -kulturen identifiziert.

Die Untersuchung hat gezeigt, dass das St. Galler Management Modell mit seiner systemischen Perspektive einen wertvollen Rahmen bietet, um die komplexen Wechselwirkungen zwischen technologischen, wirtschaftlichen und gesellschaftlichen Aspekten von Fairness zu erfassen. Die Unterscheidung zwischen Umweltsphären, Anspruchsgruppen und organisationalen Elementen ermöglicht ein differenziertes Verständnis der verschiedenen Einflussfaktoren auf faire KI-Systeme. Besonders relevant ist die Betrachtung von Fairness als strategisches Asset, was Unternehmen hilft, Fairness prominent zu positionieren und als Wettbewerbsvorteil zu nutzen. Das Leipziger Führungsmodell ergänzt diese systemische Perspektive durch seinen Fokus auf werteorientierte Führung. Die vier Dimensionen Purpose, Unternehmergeist, Verantwortung und Effektivität bieten konkrete Ansatzpunkte für Führungskräfte, um Fairness in ihren Entscheidungen zu verankern. Besonders hervorzuheben ist das Konzept des Public Value, das Fairness nicht nur als internes Firmenprinzip, sondern als gesellschaftlichen Wertbeitrag versteht. Diese normative

Dimension ist entscheidend, um Fairness nicht nur als technisches Problem, sondern als fundamentalen Aspekt verantwortungsvoller Unternehmensführung zu begreifen.

Die Integration beider Modelle ermöglicht einen holistischen Ansatz, der sowohl die strukturellen als auch die kulturellen Voraussetzungen für faire KI-Systeme adressiert. Corporate Digital Responsibility bietet dabei einen geeigneten Rahmen, um die systemische Perspektive des SGMM mit der werteorientierten Perspektive des LFM zu verbinden. Die wirtschaftsethische Dimension von CDR, zeigt, dass Fairness nicht im Widerspruch zu ökonomischen Zielen stehen muss, sondern vielmehr als Wertschöpfungsfaktor verstanden werden kann, der Vertrauen schafft, Reputation stärkt und langfristige Kundenbeziehungen fördert.

Die Fallbeispiele von Allianz und Merck illustrieren, wie Unternehmen unterschiedlicher Branchen Fairness in ihre KI-Strategien integrieren und dabei sowohl systemische als auch werteorientierte Ansätze verfolgen. Diese Beispiele zeigen, dass die Verankerung von Fairness als Wertschöpfungsfaktor kein theoretisches Konstrukt, sondern eine praktische Realität ist, die konkrete wirtschaftliche Vorteile verspricht. Gleichzeitig verdeutlichen sie die Herausforderungen, die mit der Operationalisierung von Fairness verbunden sind – von der Definition geeigneter Fairness-Kriterien über die Implementierung in komplexe Systeme bis hin zur kontinuierlichen Überwachung und Anpassung.

Die aus den theoretischen Erkenntnissen und Fallbeispielen abgeleiteten Handlungsempfehlungen bieten Führungskräften konkrete Orientierung für die Integration von Fairness in ihre Management-Praxis. Sie verdeutlichen, dass Fairness nicht als isoliertes Prinzip, sondern als integraler Bestandteil der Unternehmensstrategie betrachtet werden sollte, der geeignete organisationale Strukturen, einen differenzierten Umgang mit verschiedenen Fairness-Definitionen, eine Balance zwischen Automatisierung und menschlicher Expertise sowie ein Verständnis von Fairness als kontinuierlichem Lernprozess erfordert. Unternehmen, die Fairness in ihre Management-Modelle integrieren und dabei sowohl systemische als auch werteorientierte Perspektiven berücksichtigen, werden besser in der Lage sein, die doppelte Herausforderung von Innovation und ethischer Verantwortung zu bewältigen und nachhaltig erfolgreich zu sein. Somit führt die Orientierung an und die Befolgung von ethischen Verpflichtungen zu einer höheren Wertschöpfung. Die in diesem Beitrag entwickelten Ansätze bieten dafür einen konkreten Orientierungsrahmen, der Theorie und Praxis verbindet und so einen Beitrag zur verantwortungsvollen Gestaltung der digitalen Transformation leistet.

Becker, S. / Nemat, A. / Lucas, S. / Heinitz, R. / Klevesath, M. / Charton, J. (2022): A Code of Digital Ethics: Laying the Foundation for Digital Ethics in a Science and Technology Company, in: AI & Society, Jg 38 / Nr. 3, 2629–2639, DOI:10.1007/s00146-021-01376-w.

Brink, A. (2022): Twin Transformation Culture, in: Audit Committee Quarterly, Jg. 1, 6–9.

CDR-Initiative (2025): Corporate Digital Responsibility Initiative, URL: https://www.cdr-initiative.de (aufgerufen am:23/03/2025).

Dörr, S. (2024). Praxisleitfaden Corporate Digital Responsibility: Unternehmerische Verantwortung und Nachhaltigkeitsmanagement im Digitalzeitalter, Wiesbaden: Springer Gabler, DOI: 10.1007/978-3-662-69650-7.

Großmann, T. (2022): Fairness und Unfairness in Verhandlungen und bei der Ausführung von Verträgen im B2B-Bereich; DOI: 10.17169/refubium-37555.

Kämpf-Dern, A. (2022): St. Gallen Management Model: System-Constructionist Approach to Workspace Organisations and Management, in: Danivska, V. / Appel-Meulenbroek, R. (Hrsg.): A Handbook of Management Theories and Models for Office Environments and Services, London, New York: Routledge, 36–52.

Kirchgeorg, M. / Meynhardt, T. / Pinkwart, A. / Suchanek, A. / Zülch, H. (2019): Das Leipziger Führungsmodell: Führen und beitragen, Leipzig: HHL Academic Press.

Kovacevic, I. (2024): Integrating 'Technology with Heart': Allianz and Accenture's Insights on GenAI, URL: https://www.allianz.com/en/mediacenter/news/articles/240701-allianz-gen-ai-integrating-technology-with-heart.html (aufgerufen am: 15/05/2025).

Merck KGaA (2025): Digitale Ethik, URL: https://www.merckgroup.com/de/sustainability/business-ethics/digital-ethics.html (aufgerufen am: 15/05/2025).

Nemat, A. / Becker, S. / Lucas, S. / Thomas, S / Gadea, I. / Charton, J. (2023): The Principle-at-Risk Analysis (PaRA): Operationalising Digital Ethics by Bridging Principles and Operations of a Digital Ethics Advisory Panel, in: Minds and Machines, Jg. 33 / Nr. 1–2, 737–760, DOI:10.1007/s11023-023-09654-w.

Nuhić-Mešković, M. / Mešković, A. 2023 : Risk Management Culture, Structure, and Process – Theoretical Insights and Empirical Evidence, in: International Business Research; Jg. 16 / Nr. 10, 10–23.

Rüegg-Stürm, J. / Grand, S. (2020): Das St. Galler Management-Modell: Management in einer komplexen Welt, Bern: Haupt Verlag.

Saxena, N. A. / Huang, K. / DeFilippis, E. / Radanovic, G. / Parkes, D. C. / Liu, Y. (2020): How do Fairness Definitions Fare? Testing Public Attitudes Towards Three Algorithmic

Definitions of Fairness in Loan Allocations, in: Artificial Intelligence, Jg. 283, Artikel: 103238, DOI: 10.1016/j.artint.2020.103238.

Spiekermann, S. (2019): Digitale Ethik: Ein Wertesystem für das 21. Jahrhundert, München: Droemer Verlag.

Suchanek, A. (2020): CDR: primum non nocere, in: Bertelsmann Stiftung / Wittenberg-Zentrum für Globale Ethik (Hrsg.): Unternehmensverantwortung im digitalen Wandel, Gütersloh: Bertelsmann Stiftung, 17–22.

Teodorescu, M. / Makridis, C. (2024): Fairness in Machine Learning: Regulation or Standards?, URL: https://www.brookings.edu/articles/fairness-in-machine-learning-regulation-or-standards (aufgerufen am: 11/05/2025).

Ulrich, P. (2008): Integrative Wirtschaftsethik: Grundlagen einer lebensdienlichen Ökonomie, Bern: Haupt Verlag.

Welp, C. (2024): Digitale Vorreiterrolle – Allianz arbeitet an Versicherung der Zukunft, URL: https://www.welt.de/wirtschaft/webwelt/article253870420/Allianz-Digitale-Vorreiterrolle-Allianz-arbeitet-an-Versicherung-der-Zukunft.html (aufgerufen am: 16/05/2025).

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.

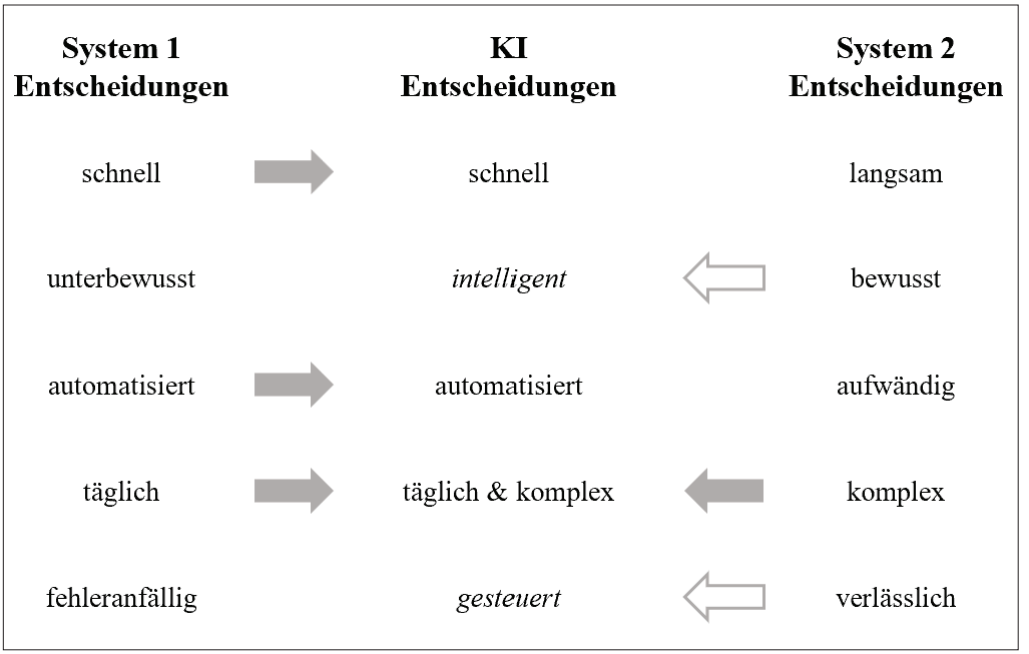


ABBILDUNG 1: ENTSCHEIDUNGSLOGIK MENSCH UND KI
(QUELLE: EIGENE DARSTELLUNG IN ANLEHUNG AN KAHNEMAN 2012)

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.

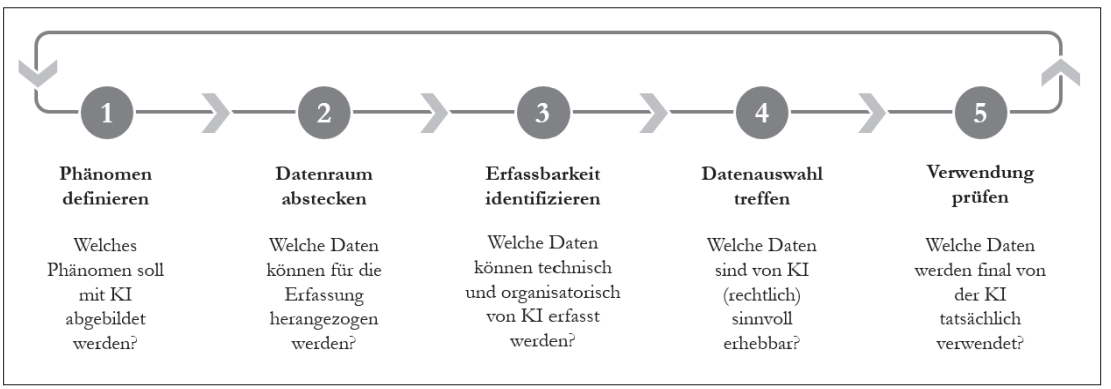


ABBILDUNG 2: REFLEXIONSLEITFADEN
(QUELLE: EIGENE DARSTELLUNG IN ANLEHUNG AN JONES 2019)

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.

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?

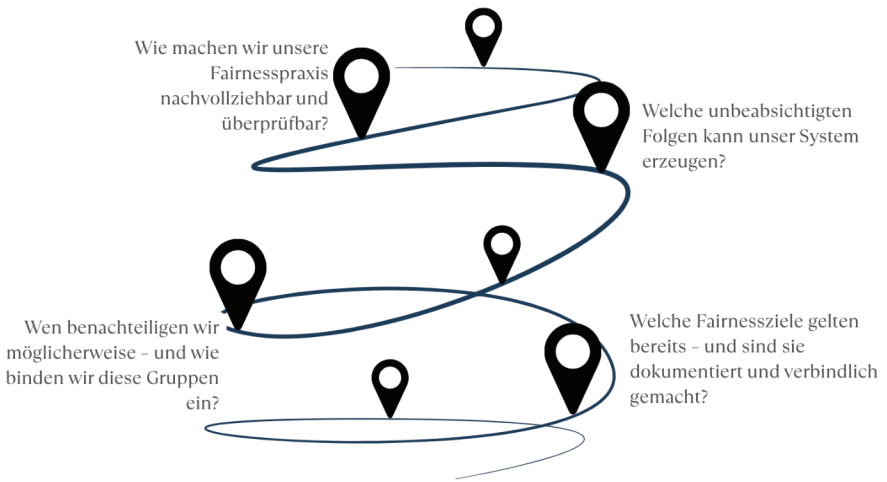


ABBILDUNG 2: FÜHRUNGSFRAGEN FÜR FAIRNESS IN KI-SYSTEMEN (OPERATIONALISIERUNG VON HUMAN IMPACT DECISION MAKING) (QUELLE: EIGENE ABBILDUNG)

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).

4.

Impulse

Technologieentwicklung und Gerechtigkeit im Zeitalter der Digitalisierung[1]

Die Diversity-Folgenabschätzung als Instrument zur Auflösung des Collingridge-Dilemmas

Emily Breuer und Olivia Hankins

1. Einleitung

Debatten um die Abwägung der Chancen und Risiken freier technologischer Entwicklung, insbesondere in Bezug auf ethische Auswirkungen von KI auf unsere Gesellschaft und unser ökonomisches Handeln, sind hochaktuell. Schon in den vergangenen Jahren wurden in Europa, als auch den USA Richtlinien für den Umgang und die Entwicklung von KI-Systemen diskutiert und letztendlich sowohl von der Europäischen Union (vgl. EU-Verordnungen 2024) als auch unter dem ehemaligen US-Präsidenten Joe Biden ausgearbeitet (vgl. O'Brian/Parvini 2024). Im US-Wahlkampf 2024 betonte vor allem Kamala Harris die Relevanz der Regulierung von technologischer Innovation, indem sie auf schwerwiegende Auswirkungen von verzerrten KI-Systemen oder deren Missbrauch Bezug nahm.

> When a woman is threatened by an abusive partner with explicit deepfake photographs, is that not existential for her? (ebd.)

[1] Workshop auf der #CDRK24 Konferenz unter dem Titel „Diversity-Folgenabschätzung: Ein Instrument, um Diversität, Inklusion und Chancengleichheit zu fördern" – geleitet von Thomas Bieler (ING Bank), Maria Hinz (Barmer) und Christian Thorun (CDR-Initiative), betreut von Emily Breuer (Universität Bayreuth) und Olivia Hankins (Universität Bayreuth). Der nachfolgende Beitrag ist im Rahmen eines Blockseminars „Ethik der Digitalisierung" an der Universität Bayreuth entstanden und wurde durch den Workshop auf der #CDRK24 Konferenz inspiriert.

Harris macht deutlich, dass KI-Systeme in ihrer Anwendung einen bedeutenden Einfluss auf die Gesellschaft haben und Fehler in ihren Algorithmen zu schwerwiegenden Folgen führen können, welche unsere demokratische Gesellschaft in Bezug auf ihre Grundpfeiler von Gleichheit und Gerechtigkeit bedrohen können. Das bedeutet, dass in frühen Entwicklungsphasen die Folgen eines solchen Algorithmus nicht adäquat erfasst worden sind. Demgegenüber stehen Haltungen, welche in ethischen Bedenken und frühzeitigen Regulierungen von KI eine Beschränkung des technologischen Fortschritts sehen, da wir die genauen Auswirkungen heute noch nicht exakt beziffern können. So hob Donald Trump nach seiner Wiederwahl die von Joe Biden beschlossenen Richtlinien zum Umgang mit KI, wie ethische Überlegungen zur Kennzeichnung von KI-Inhalten oder Sicherheitsabschätzungen der Systeme für die Gesellschaft, auf, da er diese als Hindernis für technologische Innovation betrachtete (vgl. Spiegel 2025). Dies kündigte er bereits im Wahlkampf mit den Worten: "We will repeal Joe Biden's dangerous Executive Order that hinders AI Innovation, and imposes Radical Leftwing ideas on the development of this technology" (O'Brian/Parvini 2024) an. Hinsichtlich dieser Entwicklung stellt sich auch für Unternehmen die Frage, inwiefern sie an ethischen Überlegungen und frühzeitigen Regulierungen in der Entwicklung von Technologie festhalten sollten, wenn äußere Regularien wegfallen.

In diesem Artikel beabsichtigen wir die Relevanz einer frühzeitigen Berücksichtigung ethischer Prinzipien und Folgenabschätzungen in der Technologieentwicklung zu verdeutlichen und der Frage entgegenzutreten, ob wir tatsächlich Chancen verbauen, wenn wir basierend auf noch nicht vollständig erkennbaren Auswirkungen von Technologien ethische Richtlinien für die Entwicklung setzen. Stattdessen könnten die Regulierungen sogar als Voraussetzung dafür gesehen werden, Chancen überhaupt nutzen zu können – indem sie ethische Risiken und Fehlentwicklungen minimieren. Diese Fragestellung weist auf ein Dilemma hin, welches in der Literatur als Collingridge-Dilemma bekannt ist (vgl. Collingridge 1980). Es besagt, dass die potenziellen Konsequenzen einer Implementierung der Technologie in einem frühen Entwicklungsstadium in ihrem vollen Umfang noch unbekannt sind und sich erst nach der Implementierung zeigen und daher schwer zu berücksichtigen sind. Zeigen sich die Folgen erst nach der Implementierung, so ist ein Gegensteuern aufgrund der bereits vielseitigen Verwurzelung der Technologie schwierig und kostspielig, woraus sich ein Dilemma ergibt. Obwohl die Unsicherheit über die Zukunft das Dilemma schwer aufzulösen scheint, hat ein Versuch höchste Relevanz, womit sich die Frage nach einer Möglichkeit der

Auflösung des Dilemmas ergibt, d.h. wie man ethische Regulationen schon im Entwicklungsprozess einbindet, um schwerwiegende Folgen zu adressieren, die man noch nicht genau kennt.

In diesem Artikel möchten wir das Instrument der Diversity-Folgenabschätzung, welches von der Corporate Digital Responsibility Initiative für Unternehmen entwickelt wurde, als Versuch betrachten, das Collingridge-Dilemma aufzulösen und es in Bezug auf sein philosophisches Fundament analysieren. Die Corporate Digital Responsibility (CDR) Initiative unterstützt Unternehmen, Verantwortung für eine ethisch fundierte Umsetzung der Digitalisierung zu übernehmen, wobei insbesondere der Mensch in den Vordergrund gestellt wird (vgl. CDR-Initiative 2024). Die Diversity-Folgenabschätzung, welche von ihr als Instrument für unternehmensinterne Digitalisierungsprojekte entworfen wurde, soll ethische Überlegungen zu Inklusion und Diversität schon in den Entwicklungsprozess implementieren, um negative Auswirkungen wie einer Diskriminierung von Mitarbeitenden über inadäquate Angepasstheit an Fähigkeiten im Umgang mit digitalen Systemen, vorzubeugen. Im Folgenden werden wir die Diversity-Folgenabschätzung bezüglich ihres Potenzials das Collingridge Dilemma zu adressieren betrachtet. Dabei wird sich trotz einiger Unzulänglichkeiten nicht nur zeigen, dass sie das Dilemma in relevanter Weise adressieren kann, sondern auch dass sie Unternehmen eine gute Grundlage bietet ihrer Verantwortung und eigenem Interesse in der Auflösung des Dilemmas nachzukommen. Ihre gesellschaftliche und ethische Relevanz zeigt sich dabei vor allem auch in ihrer Anwendbarkeit auf Szenarien, wo keine politischen Regularien für Unternehmen seitens der Politik getroffen werden.

2. Die Relevanz der Berücksichtigung ethischer Prinzipien in der Entwicklung von Technologie

2.1 Die Perspektive eines Laissez-faire-Ansatzes und die Regulierung der Entwicklung und des Einsatzes von Künstlicher Intelligenz

Im Zeichen der Debatte um die (ethische) Notwendigkeit von Regularien in der KI-Entwicklung bestehen unterschiedliche Positionen. Dabei werden Regularien einerseits als ein Hemmnis für die Entwicklung fortschrittlicher Technologie gesehen und ethischen Bedenken in der Entwicklung wenig Raum geboten, woraus eine Laissez-Faire-Haltung erwächst, bei welcher technologischer Fortschritt weitgehend ohne regulatorische Eingriffe gefördert wird und damit

keine ethischen Richtlinien für die Unternehmen vorgegeben werden, sodass potenzielle aber noch nicht genau absehbare gesellschaftliche Konsequenzen nicht zwangsläufig Berücksichtigung finden, bzw. erst nach der Implementierung untersucht werden.

Auf der anderen Seite stehen Positionen, welche eine hohe Relevanz in der Vorgabe einheitlicher ethischer Maßstäbe, wie Transparenzanforderungen oder Tests vor der Implementierung, sehen, um ethischen Konsequenzen für die Gesellschaft aus der Entwicklung von KI-Systemen vorzubeugen und Unternehmen Anreize zu setzen ethische Betrachtungen in ihre Entwicklungen einzubeziehen. Als Beispiel lässt sich hier der „AI-Act" der Europäischen Union anführen, welcher Regulierungen für künstliche Intelligenz umfasst und (ethische) Risikoabwägungen trifft (vgl. Europäisches Parlament 2023). Man versucht Richtlinien für Folgen zu implementieren, deren Ausmaß noch nicht feststeht, um Schaden von der Gesellschaft abzuwenden.

Aus der Unternehmensperspektive erscheint ein Laissez-Faire-Ansatz zunächst attraktiver, da er zulässt, dass sich Technologie frei entwickeln kann, alle Möglichkeiten offenstehen und geringere Kosten in der Entwicklung anfallen, wenn nicht zusätzlich ethische Überlegungen und Richtlinien zu berücksichtigen sind und Ressourcen nicht durch etwaige zukünftige Folgen, welche heute noch nicht im Detail bestimmbar sind, gebunden sind. Es erscheint, als könnte er dazu beitragen, dass bahnbrechende Technologien schneller zur Marktreife gelangen und Wettbewerb sowie wirtschaftliches Wachstum stimuliert werden. Dies bedeutet aber auch bedeutende Risiken in unvorhergesehenen negativen Auswirkungen, da negative Auswirkungen für die Gesellschaft von einem Unternehmen unter dem Laissez-Faire-Ansatz nicht zwangsläufig internalisiert werden, da sie nicht mit übergeordneten Richtlinien an die ethischen Konsequenzen ihrer Entwicklung gebunden sind. Mit der Abschaffung einheitlicher gesetzlicher Richtlinien zur Entwicklung und Einsatz von KI durch Donald Trump stehen Unternehmen in den USA einer Laissez-Faire-Situation gegenüber (vgl. Spiegel 2025), woraus die Frage erwächst, ob dennoch eine Betrachtung ethischer Maßstäbe in der Entwicklung ihrer Technologie eingebunden bleiben sollten.

Im Folgenden werden wir vor diesem Hintergrund erläutern, weshalb insbesondere für die Unternehmen auch unabhängig von gesetzlichen Vorgaben eine Relevanz in der frühzeitigen Berücksichtigung ethischer Überlegungen und der gesellschaftlichen Folgen von Technologieentwicklungen besteht. Es wird sich zeigen, dass ethische Überlegungen in ihrem eigenen wirtschaftlichen Interesse liegen und eine Laissez-Faire-Situation sie nicht von ihrer gesellschaftlichen Verantwortung in der Technologieentwicklung entbindet.

Unsere Gesellschaft basiert in einem hohen Maße auf ethischen Prinzipien, welche unsere politische Organisation, unser Wirtschaften und das Zusammenleben bestimmen. Es bestehen verschiedene Vorstellungen von einer gerechten und moralischen Gesellschaft, wobei wir in diesem Zusammenhang vor allem auf die Vertragstheorie John Rawls' verweisen wollen. Rawls beschreibt die Gesellschaft als kooperatives System, dessen Gerechtigkeitsprinzipien auf der Grundlage eines fairen Entscheidungsprozesses basieren (vgl. Rawls 1999). Er stellt eine hypothetische Ausgangssituation dar, in der alle Mitglieder der Gesellschaft über die grundlegenden Prinzipien sozialer Gerechtigkeit entscheiden, dabei jedoch nicht wissen, welche soziale Position sie später einnehmen werden. Da niemand weiß, ob er oder sie reich oder arm, privilegiert oder benachteiligt sein wird, würden rationale Individuen Prinzipien wählen, die für alle fair sind (vgl. ebd.: 15–19). Unterschiedliche körperliche und soziale Merkmale sind dabei moralisch willkürlich, weshalb kein darauf basierender Anspruch auf mehr Rechte besteht (vgl. Wenar 2021). Dies impliziert ein Prinzip der Gleichheit und Nichtdiskriminierung. Ethische Prinzipien wie Gleichheit und Gerechtigkeit sind somit nicht nur abstrakte Werte, sondern in der gesellschaftlichen Übereinkunft unseres gemeinsamen Wirtschaftens und Zusammenlebens verankert. Auch wenn diese Überlegungen in einem vorgesellschaftlichen Zustand hypothetisch sind und keine tatsächliche vertragliche Vereinbarung besteht, zeigen sie auf, dass Gerechtigkeitsprinzipien rational fundiert sind und aus ihr für die Mitglieder eine Verantwortung ihrer Bewahrung und Verwirklichung erwächst. Dies ist keine rein politische und institutionelle Verantwortung, welche allein die Wahrung von Rechten und einer staatlichen Struktur betrifft, sondern betrifft jede unserer Handlungen, welche eine Wirkung auf gesellschaftliche Zusammenhänge hat, unter anderem auch die Entwicklung von Technologie.

Es besteht ein Bewusstsein darüber, dass neue technologische Entwicklungen unsere Gesellschaft prägen und verändern, woraus auch ein verantwortungsvoller Umgang hinsichtlich ihrer Berührung ethischer Prinzipien erwächst. Dies kann derzeit vor allem vor dem Hintergrund der Digitalisierung erkannt werden. So findet beispielsweise ein großer Teil unserer Kommunikation mittlerweile in digitalen Räumen statt, die häufig als frei zugängliche und gleichberechtigte Plattformen des Meinungsaustauschs illustriert werden. Dennoch sind Plattformen, wie Facebook oder X (ehemals Twitter) ebenso über technische Strukturen wie Algorithmen geprägt, wodurch Aufmerksamkeit und Reichweite gelenkt werden und nicht allein in der Hand des Nutzers liegen.

Die neue Technologie formt damit die Struktur unserer gesellschaftlichen Kommunikation, woraus aber auch eine Verantwortung in der Entwicklung erwächst, ethische Prinzipien, auf welchen unsere Gesellschaft baut, auch im digitalen Raum zu erhalten und diese in die Technologie zu implementieren, damit beispielsweise Diskriminierungen im digitalen Raum ebenso sozial sanktioniert werden und sich nicht strukturell verfestigen (vgl. Seeliger/Sevignani 2022: 3–16).

Ein weiteres Beispiel besteht in KI-gestützten Bewerbermanagementsystemen, welche Bewerbungsverfahren für eine offene Stelle optimieren sollen. Auch hier gilt es in ihrer Entwicklung ethische Prinzipien, wie der Inklusion und Nicht-Diskriminierung, hinsichtlich des Geschlechts, der Herkunft oder einer Behinderung zu berücksichtigen, da sie auf unsere gesellschaftlichen Strukturen, wie die Zusammensetzung des Personals und Bewerbungschancen wirken. KI-Systeme basieren dabei auf menschlichen Daten, wodurch die Gefahr besteht, dass bestehende strukturelle soziale oder politische Voreingenommenheit, die in der Gesellschaft bestehen, in maschinell lernende Algorithmen als Muster wahrgenommenen und verfestigt werden, sodass sich Entscheidungsprozesse darauf berufen (vgl. Ruttkamp-Bloem 2023: 20–23). Wenn der Fall besteht, dass diese Systeme mit voreingenommenen Daten trainiert wurden, sodass z. B. in der Vergangenheit überwiegend männliche Bewerber bevorzugt wurden, woraufhin die entwickelte KI Frauen oder andere Gruppen systematisch benachteiligt, so betrifft dies die Strukturen unserer Gesellschaft und in einem größeren Kontext die strukturelle Verankerung der ethischen Prinzipien. Das heißt, dass die Technologie nicht nur Entscheidungskriterien für die Vergabe von Jobs berührt, sondern über mögliche Diskriminierungen auch fundamentale gesellschaftliche Prinzipien der Gleichberechtigung oder Inklusion verletzen kann. Dies wird bei einer langfristigen Implementierung nicht nur einzelne Jobvergaben beeinflussen, sondern auch die Umsetzung ethischer Prinzipien prägen und damit die strukturelle Art und Weise, wie Entscheidungen in unserer Gesellschaft getroffen werden. Die Relevanz der Implementierung ethischer Überlegungen und Folgenabschätzungen bei der Entwicklung von Technologie besteht somit darin, dass sie bedeutende Einflüsse auf unsere gesellschaftlichen Strukturen und die Verwirklichung unserer ethischen Prinzipien haben, welche auf eine gerechte Gesellschaft abzielen.

Die Verantwortung, die mit der Wirkung von Technologieentwicklungen auf die gesellschaftliche Struktur einhergeht, liegt dabei jedoch nicht allein in der Vorgabe von Regularien seitens politischer Institutionen, sondern auch bei jedem Unternehmen (vgl. Strasser/Niedermayer 2021),

da Unternehmen in einem hohen Maße mit der Gesellschaft verflochten sind und in einem Wechselspiel stehen. Ihre Produkte und Dienstleistungen, d.h. auch ihre technologischen Innovationen stellen Lösungen und Angebote für die Bewältigung gesellschaftlicher Prozesse und Bedürfnissen dar wie z. B. die Informationsbeschaffung mit Hilfe intelligenter Suchmaschinen oder KI-basierter Auswahlsysteme für Jobs oder Studienplätze. Diese betreffen mitunter bedeutende und lebensverändernde Angelegenheiten, da Entscheidungsprozesse der Systeme uns direkt betreffen oder wir aufbauend auf unserem Informationsstand politische, ökonomische, aber auch individuelle Entscheidungen treffen, wodurch die Ausgestaltung der Technologie durch die Entwickler relevanten Einfluss auf das Leben ihrer Nutzer und darüber hinaus langfristig auf die gesellschaftliche Struktur hat.

2.3 Unternehmensethik im Kontext von KI

Eine Implementierung ethischer Prinzipien in der Entwicklung von KI ist jedoch nicht nur aufgrund der gesellschaftlichen Verantwortungsbeziehung eines Unternehmens relevant, sondern ist auch in ihrem eigenen Interesse begründet. Auch wenn eine Laissez-Faire-Haltung hinsichtlich Innovationen vielversprechender erscheinen mag, wird ein Unternehmen langfristig von einer frühzeitigen Implementierung ethischer Überlegungen profitieren. Die Einbindung ethischer Überlegungen in die Entwicklung von Technologie ist auch eine Strategie der Risikominimierung von gesellschaftlichen Folgen und damit auch für die Nutzer. Dazu gehören nicht nur rechtliche Aspekte, wie etwa die Haftung für diskriminierende oder fehlerhafte KI-Entscheidungen, sondern auch wirtschaftliche Risiken. Investitionen in Technologien, die sich später als ethisch problematisch und gesellschaftlich inakzeptabel erweisen, können zu finanziellen Verlusten führen und das Vertrauen der Öffentlichkeit sowie die Reputation eines Unternehmens erheblich schädigen. Auch in Bezug auf die Marktakzeptanz spielt die frühzeitige Einbindung ethischer Überlegungen eine entscheidende Rolle, da Verbraucher und Unternehmen zunehmend Wert auf verantwortungsbewusste Technologien legen, die Transparenz, Fairness und gesellschaftliche Werte berücksichtigen, woraus sich eine sinkende Wettbewerbsfähigkeit ergeben kann. So zeigen Anissa Dakhli's Untersuchungen zur Beziehung zwischen der sozialen Verantwortung und der finanziellen Leistungsfähigkeit von Unternehmen, dass ein positiver Einfluss auf die finanzielle Leistung eines Unternehmens besteht, wenn ethisches Verhalten und Transparenz in seinen Aktivitäten berücksichtigt werden (vgl. Dakhli 2021). Die Berücksichtigung ethischer Prinzipien ist

dabei jedoch nicht nur eine Reaktion auf gesellschaftliche Entwicklungen, sondern auch in der Anreizstruktur eines Unternehmens verankert. Eine diverse und inklusive Unternehmenskultur trägt dazu bei, vielfältige Perspektiven und Innovationsideen zu fördern, was die Wettbewerbsfähigkeit langfristig stärken kann. Auch wenn Gerechtigkeitsprinzipien in erster Hinsicht auch als Einschnitt in die Wirtschaftlichkeit eines Unternehmens begriffen werden können, z. B. über Quoten zur Einstellung bestimmter Personengruppen. So führt, wie Scott E. Page untersucht hat, ein diverses Umfeld langfristig zu einer besseren Anpassungsfähigkeit, kreativeren Lösungsansätzen und einer erhöhten Innovationskraft (vgl. Page 2007).

Durch die ethische Fundierung unserer Gesellschaft wird die Entwicklung neuer Technologie, welche im Kontext der Digitalisierung gesellschaftliche Strukturen neugestaltet, an die Berücksichtigung ethischer Folgen gebunden sein und die Entwicklung von z. B. ethischen KI-Systemen einen Anspruch darstellen. Damit wird auch der wirtschaftliche Erfolg der Unternehmen und ihre Wirtschaftsweise im Kontext der Digitalisierung von ihrer Implementierung ethischer Folgenabschätzungen abhängen (vgl. Lauer 2021: 21–23). Dies stellt jedoch eine bedeutende Herausforderung dar, da die genauen Auswirkungen einer Technologie sich erst in ihrer zukünftigen Implementierung zeigen und im Entwicklungsprozess, zwar teilweise kalkuliert, aber schwer reguliert werden können. Dies ist vor allem im Fall der KI-Systeme ersichtlich, da ihre Entscheidungsprozesse zwar auf ausgewählten Trainingsdaten basieren, sie jedoch kontinuierlich über die Interaktion mit weiteren Daten hinzulernen, was die Abschätzung ihrer Entwicklung und Ausgestaltung und damit auch ihrer gesellschaftlichen Auswirkungen erschwert. Wir kennen die genauen Folgen damit erst, wenn sie entstehen, was es schwer macht, Regulierungen präzise genug zu gestalten, sodass sie die Folgen abfangen, bzw. ethische Bedenken adäquat auf die Zukunft auszurichten. Damit stellt sich die Frage, wie wir heute alle ethischen Folgen der Technologie abschätzen können, die erst morgen sichtbar sein werden. Kann dieses Dilemma über die Implementierung ethischer Überlegungen in der Technologie gelöst werden?

Die Erläuterungen dieses Kapitels haben gezeigt, dass auch wenn seitens der Politik eine Laissez-Faire-Haltung bezüglich der ethischen Regularien besteht, für Unternehmen eine Relevanz in der frühzeitigen Berücksichtigung ethischer Auswirkungen ihrer Technologieentwicklung besteht, welche nicht nur in ihrer gesellschaftlichen Verantwortung, sondern auch in einem ökonomischen Eigeninteresse begründet liegt. Die unternehmerische Praxis eines regulatorischen Ansatzes über die Implementierung von ethischen Prinzipien in die Entwicklung von Technologie

beinhaltet dennoch Unsicherheiten hinsichtlich ihrer Umsetzung, da es in einem frühen Stadium der Entwicklung schwer ist die Folgen abzuschätzen, während sie im Nachhinein eine zu starke Verankerung und auch schon Auswirkungen aufweisen. Dieses so genannte Collingridge-Dilemma wird im Folgenden erläutert."

3. Das Collingridge-Dilemma

Das Collingridge-Dilemma, welches auf David Collingridge zurück geht, ist vor allem im Bereich der Folgen und Risikoabschätzung neuer Entwicklungen bedeutsam, insbesondere in Bezug auf deren Wirkung hinsichtlich der Ausgestaltung und Beförderung einer gerechten und inklusiven Gesellschaft (vgl. Collingridge 1980). Dieses Dilemma besteht darin, dass früh im Entwicklungsprozess der Technologie die potenziellen Konsequenzen dieser Technologie noch nicht bekannt sind, sondern die Effekte erst zu einem späteren Zeitpunkt nach ihrer Implementierung sichtbar werden. Zu diesem Zeitpunkt ist es allerdings schwierig, die Technologie noch zu regulieren, da Prozesse schon fest verankert sind. Kurz gesagt ist dieses Dilemma ein Problem der Zeit und Technologieentwicklung (vgl. Demos Helsinki 2022). Im Collingridge-Dilemma gibt es zwei Bedingungen, welche bestehen müssen, um die unerwünschten Konsequenzen einer Technologie vermeiden zu können. Zum einen muss Wissen darüber bestehen, dass eine Technologie schwerwiegende Konsequenzen hat oder in Zukunft haben wird. Zweitens muss es möglich sein, die Technologie im späteren zu verändern, um diese Konsequenzen zu beheben. In der Praxis bleiben diese Bedingungen jedoch häufig unerfüllt, weshalb wir einem Dilemma gegenüberstehen.

Ein konkretes Beispiel, das dieses Dilemma eindeutig illustriert, besteht in der Entwicklung des Autos. Im Jahr 1908 waren Veränderungen im Design und der Benutzung von Autos einfach umzusetzen, wobei Prognosen zu späteren sozialen Konsequenzen noch schwer möglich waren. Die heutzutage auftretenden Folgen wie eine erhebliche Luftverschmutzung sowie soziale Veränderungen wie die Prägung unseres Mobilitätsverhalten, der Stadtarchitektur und einer Beförderung sozialer Ungleichheiten waren im Detail nicht vorherzusagen. Heute sind diese Folgen fast jeden Tag sichtbar und effektive Änderungen schwer durchzuführen, da sie sich fest in unserem gesellschaftlichen Zusammenleben verankert haben. Dieses Dilemma besteht auch bezüglich ei-

ner Folgenabschätzung zu ethischen Konsequenzen der Digitalisierung. Wir kennen beispielsweise heute die Auswirkungen einer zunehmenden Etablierung von KI-Assistenten auf unsere Gesellschaft nicht im Detail, da diese erst sichtbar werden, nachdem die Technologie schon in der Gesellschaft und Wirtschaft verankert ist. Trotz frühen Bemühungen zur Regulierung von KI-Systemen bleibt eine effektive Regulierung kompliziert, weil man in Anbetracht der Unsicherheit der zukünftigen gesellschaftlichen Struktur und der Etablierung der Systeme nicht weiß, wo der Fokus der Regulierung liegen sollte. Dieser kann lediglich über heutige Normen und Wertvorstellungen abgeschätzt werden. Das Dilemma von Collingridge kann damit auch folgendermaßen formuliert werden:

> Either one attempts to control technology by early prediction of consequences, or one attempts to control technology by reacting to consequences as they unfold. If the former, then the attempt fails because of predictive unreliability. If the latter, then the attempt fails because of developed rigidity. Therefore, either the attempt to control technology fails because of predictive unreliability, or the attempt to control technology fails because of developed rigidity (Croy 1996: 2).

Das Collingridge-Dilemma betrifft damit den Zeitverlauf und den Entwicklungsprozess. In der Literatur bestehen unterschiedliche Lösungsansätze des Dilemmas. Unter anderen von Daniel Tigard, welcher diskutiert, inwiefern eine Brücke zwischen den beiden Zeitpunkten in der Technologieentwicklung und ihren Wissensständen geschlagen werden kann, sodass Folgen in frühen Stadien regulierbar sind und das Dilemma gelöst werden kann (vgl. Tigard 2022: 12-19). Er verweist unter der Bezeichnung "embedded ethics" auf flexible Ansätze, welche eine Vielzahl an Perspektiven in den Entwicklungsprozess einbeziehen, um mögliche Konsequenzen über die darin enthaltenen unterschiedlichen Erfahrungen und Bedürfnisse zu erkennen (vgl. ebd.: 7–20). Dabei soll insbesondere ein Bewusstsein für ethische und soziale Wirkungen von Technologie eingebettet werden und Sensibilität gegenüber Fehlern und Verzerrungen in Daten und Technologie auf Grundlage eines Abgleichs mit ethischen Prinzipien entstehen (vgl. ebd.: 17–27). Mit der Berücksichtigung ethischer Überlegungen und ihrer Einbettung in den frühen Entwicklungsprozess tut sich dabei jedoch eine weitere Dimension des Dilemmas auf und zeigt, dass diese Brücke zwischen beiden Zeitpunkten des Dilemmas auch mit einer solchen Einbettung schwer überwindbar scheint. Dies ist der Fall, weil technologische Entwicklungen

gesellschaftliche Umstände beeinflussen und dabei auch Normen und Werte formen. Wenn wir jedoch unsere heutigen Wertmaßstäbe anlegen, berücksichtigen wir nicht jene, die in der Zukunft gelten werden (vgl. Kudina/Verbeek 2019).

Kudina und Verbeek, adressieren dies in Abgrenzung zur wirtschaftlichen Variante des Dilemmas als ethische Version, welche darauf basiert, dass die Technologie oftmals auch unsere herangezogenen Werte verändert (vgl. ebd.). Dabei besteht das Problem, dass ethische Regularien entweder zu früh, d.h. bevor klar ist, wie die Technologie Bewertungsmaßstäbe beeinflussen wird, eine Bewertung der Technologie vornehmen oder zu spät, wo die ethischen Auswirkungen zwar bekannt sind, die Technologie aber bereits etabliert und kaum noch veränderbar ist (vgl. ebd.: 2–5). Laut den Autoren gibt es zwei Wege dieses Dilemma zu adressieren. Zum einen über einen techno-moralischen Weg, der versucht, die ethischen Aspekte einer Technologie vorherzusagen und zu antizipieren (vgl. ebd.: 292–295). Dabei ist dieser Weg kaum zuverlässig, weil er keine Dynamiken zwischen Technologie und Moralität adressiert, sondern nur ihre Implikationen. Zum anderen besteht der sozio-technische Weg, welcher Voraussagen über die Zukunft als spekulativ zurückweist und stattdessen Technologieentwicklungen als soziale Experimente auffasst, welche ethische Überlegungen in ihrer Durchführung erfordern. Er versucht damit, den Prozess technologischer Entwicklung zu regulieren (vgl. ebd.: 3). Die Autoren selbst plädieren, basierend auf ihrer These, dass sich normative Werte zusammen mit neuer Technologie verändert haben, für ein Vorgehen, das sich auf die Dynamiken der Interaktion zwischen Mensch und Technologie fokussiert und die ethischen Wechselwirkungen analysiert (vgl. ebd.: 1–3).

> [W]hen technologies are used, they typically do not play a role as technological "objects" in interaction with human "subjects"; rather, they are "mediators" of the relation between users and their environment (ebd.: 7).

Wenn man die Entwicklung Technologies somit als ein dynamisches System versteht, gibt es mehr Möglichkeiten die Auswirkungen der Technologie auf die Gesellschaft zu verstehen. Nicht nur die Gesellschaft, sondern auch die Technologieentwicklung, sind kein stagnierendes Objekt, sondern verändern sie sich stetig und beeinflussen sich gleichzeitig.

Die Relevanz der Auflösung des Collingridge-Dilemmas ist eng mit der Relevanz des Einbezugs ethischer Überlegungen in die Entwicklung von Technologie verknüpft. Unternehmen haben ein Interesse daran, dass ihre Technologie zuverlässig funktioniert und in Zukunft zu keinen

negativen Auswirkungen führt, welche die Effizienz und Anwendbarkeit der Technologie gefährden, als auch daran, dass sich keine negativen Effekte auf die Gesellschaft ergeben, sodass sie ihrer Verantwortung nachkommen. Das Collingridge-Dilemma fasst dabei die Problematik zusammen, vor welcher Entwickler in der Entwicklung stehen. In einem frühen Stadium, vor der Implementierung sind Folgen noch nicht sichtbar und sind schwer zu regulieren, da sie sich erst zeigen, wenn die Technologie implementiert und eine Korrektur schwierig oder sehr kostspielig wird. Es liegt somit im Interesse eines Unternehmens, dieses Dilemma zu lösen, um möglichst geringe Risiken einzugehen und schwerwiegende Folgen zu vermeiden.

Im Folgenden wollen wir einen Ansatz vorstellen, welcher genau diesen Anspruch verfolgt und sich zur Aufgabe macht ethische Überlegungen zu Folgen technologischer Entwicklungen auf die Diversität in den Entwicklungsprozess zu implementieren, um diese zu regulieren und im späteren nicht auf schwerwiegende Konsequenzen für die Gesellschaft als auch für das Unternehmen zu stoßen. Wir werden diesen, als einen Versuch der Lösung des Collingridge-Dilemmas diskutieren und aufzeigen, inwiefern er das Dilemma auflösen kann und wo Verbesserungspotenzial oder auch Schwierigkeiten bestehen.

4. Die Diversity-Folgenabschätzung

4.1 Die Diversity-Folgenabschätzung der Corporate Digital Responsibility Initiative

Die Corporate Digital Responsibility Initiative ist eine Initiative des Bundesministeriums für Umwelt, Naturschutz, nukleare Sicherheit und Verbraucherschutz und bietet Unterstützung sowie ein Austauschforum für eine verantwortungsvolle Umsetzung der digitalen Transformation in Unternehmen. Dabei stehen neben der digitalen Veränderung insbesondere die sozialen und ethischen Auswirkungen dieses Wandels im Vordergrund und sollen Berücksichtigung finden. Im Jahr 2024 veröffentlichte die CDR-Initiative ein Instrument zur Diversity-Folgenabschätzung, für die Förderung von Diversität, Inklusion und Chancengleichheit im Prozess der Digitalisierung, welche Unternehmen als Leitfaden dienen soll (vgl. CDR-Initiative 2024). Dabei besteht die Kernidee darin, dass schon im Entwicklungsprozess digitaler Produkte und unternehmensinternen Dienstleistungen eine Vielzahl an Perspektiven und ethischen Überlegungen zur Betroffenheit implementiert werden, um Risiken nach der Implementierung

schon in der Entwicklung zu erkennen und besser abschätzen zu können (vgl. ebd.: 8–11). Darin lässt sich ein Kalkül des Versuchs einer Auflösung des Collingridge-Dilemmas, erkennen, da versucht wird Folgen, die eigentlich erst nach der Implementierung sichtbar werden und dann schwer zu verbessern sind über eine frühzeitige Berücksichtigung ethischer Standards als auch den Einbezug vieler Perspektiven, Betroffenheiten und Erfahrungen abzuschätzen. Unsere These lautet somit, dass die Diversity-Folgenabschätzung versucht, das Collingridge-Dilemma aufzulösen. Anlässlich einer genaueren Betrachtung der Diversity-Folgenabschätzung soll in diesem Teil zunächst das von der CDR-Initiative herausgegebene Papier zur Diversity-Folgenabschätzung dargestellt werden.

Laut der CDR-Initiative besteht eine gesetzliche, gesellschaftliche und wirtschaftliche Verantwortung, sich mit ethischen Konsequenzen in der Technologieentwicklung zu beschäftigen (vgl. ebd.: 9). Die Diversity-Folgenabschätzung ist dabei als Instrument geeignet, da sie einen einfachen Prozess zur Sensibilisierung für die Thematik der Inklusion, eine Identifikation von Verbesserungspotenzialen sowie einen bedeutsamen Beitrag in der Einbindung ethischer Überlegungen zum wirtschaftlichen Erfolg eines Unternehmens bietet (vgl. ebd.). Adressiert an potenzielle unternehmerische Nutzer formulieren dies folgendermaßen:

> Die Diversity-Folgenabschätzung ermöglicht es Ihnen, im Rahmen einer Vorabkontrolle zu überprüfen, welche Auswirkungen die Einführung von neuen digitalen Lösungen auf Ihre Mitarbeitenden, Kund*innen und andere Stakeholder hat (CDR-Initiative 2024: 10).

Die Diversity-Folgenabschätzung wird in drei Phasen gegliedert. Die erste Phase, die Vorbereitungsphase, ist dabei in vier Schritte unterteilt, wobei insbesondere die Festlegung des Anwendungsfalls und der zu untersuchenden Diversitäts-Dimensionen sowie die Identifikation der relevanten Stakeholder zu nennen sind. Um einen spezifischen Anwendungsfall wie z. B. die Entwicklung einer Mitarbeiterplattform, und die damit verbundenen Auswirkungen in Bezug auf Diversität und Inklusion klar abzugrenzen, gibt die Initiative den Unternehmen einen Fragekatalog an die Hand, welcher spezifische Fragen nach der Zielsetzung oder erhofften Verbesserungen adressiert (vgl. ebd.: 15–21). In einem weiteren Schritt werden dann die Diversitätsdimensionen, welche im konkreten Fall untersucht werden sollen, festgelegt, d.h. es werden Dimensionen, wie Alter, Geschlecht oder die Vereinbarkeit von Beruf und Familie

festgelegt, bezüglich welcher man eine Folgenabschätzung durchführen möchte, definiert. Wesentlich für die erfolgreiche Durchführung der Folgenabschätzung ist dabei, einen konkreten Anwendungsfall zu behandeln und zunächst nur einige und nicht zu hohe Ziele zu stecken, um diese effektiv zu bearbeiten, auch wenn im Prinzip noch viel mehr Dimensionen des Falls betrachtet, werden könnten (vgl. ebd.: 15). Im Anschluss daran gilt es, die relevanten Stakeholder in Bezug auf ihre Betroffenheit im konkreten Anwendungsfall zu identifizieren. Mit der Bezeichnung der relevanten Stakeholder ist daran zu erinnern, dass die Diversity-Folgenabschätzung nur in einem interdisziplinären Team erfolgreich ist, wobei relevante Stakeholder zum Beispiel Mitarbeiter des Bereichs Human Ressource (HR), Diversitätsbeauftragte, Menschen aus den Dimensionen-Zielgruppen oder der Betriebsrat sein können. Damit alle Stakeholder an einem Workshop beteiligt werden können, gilt es seitens des Unternehmens eine geeignete Agenda und ein Format festzulegen.

Die zweite Phase beinhaltet zwei Schritte, die sich mit dem eigentlichen Workshop zur Folgenabschätzung bei der Entwicklung von digitalen Tools beschäftigen (vgl. ebd.: 22–24). Laut dem Papier sollte der Workshop die Relevanz der Thematik praxisnah abbilden und den konkreten Anwendungsfall in der Diskussion mit allen Stakeholdern hinsichtlich der Folgen für die Diversität analysieren (vgl. ebd.). Weiterhin sollten die zentralen Ergebnisse nach dem Workshop gesammelt und zusammengefasst werden, wobei nicht nur die (kritischen) zentralen Ergebnisse, sondern auch die Zielsetzung des Workshops und priorisierte Verbesserungsmaßnahmen berücksichtigt werden sollten.

In der dritten Phase, der Umsetzungs- und Reflexionsphase, sollen die Erfahrungen des Workshops reflektiert und die entwickelten Verbesserungsmaßnahmen implementiert werden (vgl. ebd.: 25–26). Weil jedes Unternehmen und ihr jeweiliger Anwendungsfall und die Verbesserungsmaßnahmen unterschiedlich sind, gibt es keine konkreten Empfehlungen seitens der Initiative, wie die Verbesserungsmaßnahmen implementiert werden sollten. Nach einiger Zeit sollten die Unternehmen jedoch erneut nachfragen, inwieweit die Verbesserungsmaßnahmen geholfen haben und die Ergebnisse und Erfahrungen mit der Diversity-Folgenabschätzung kritisch reflektieren, wobei die Stakeholder auch in Bezug auf ein Feedback eingebunden werden sollten (vgl. ebd.: 25 f.). Die Methode einer Diversity-Folgenabschätzung kann sich so immer weiterentwickeln, wozu zudem eine kurze Checkliste nach allen drei Phasen besteht. Die CDR-Initiative beschreibt auch zwei Beispiele Anwendungsfälle von der

ING und der Barmer. Um besser zu verstehen, wie diese Diversity-Folgenabschätzung aufgebaut ist, betrachten wir das Fallbeispiel der ING (vgl. ebd.: 18). Das Unternehmen nutzt ein internes digitales Tool für die Abrufung eines Gesundheitsbudgets für Mitarbeiter, welches Informationen und Möglichkeiten zu angebotenen Gesundheitsmaßnahmen bereitstellt. Im ersten Teil der Diversity-Folgenabschätzung konnten die relevanten Stakeholder identifiziert werden, die in der Nutzung des Tools wegen ihres Alters, Geschlechts oder anderen Diversitätsdimensionen benachteiligt werden könnten. Zum Beispiel waren einige digitale Icons in ihrer Bedeutung nicht für jeden selbstverständlich oder einige Schriftarten für einige schwer zu lesen (vgl. ebd.). Damit wurden Probleme identifiziert, die ohne einen konkreten Einbezug vielfältiger Perspektiven und Bedürfnisse unerkannt geblieben wären, da sie für Unbetroffene nicht ersichtlich waren. Nach dem Workshop standen damit konkrete Verbesserungen des digitalen Tools, hinsichtlich seiner gleichberechtigten Nutzbarkeit bereit. Die ING berichtet diesbezüglich:

> Die Diversity-Folgenabschätzung hat sich dabei als Methode bewährt und uns – entgegen unserer ursprünglichen Einschätzung – Verbesserungsbedarf bei der Gestaltung des Tools und der enthaltenen Texte aufgezeigt (ebd.).

Dieses Beispiel zeigt, wie wichtig Diversity-Folgenabschätzung in der Technologieentwicklung ist, da ohne eine spezifische Betrachtung von Perspektiven und Betroffenheiten in Bezug auf die Diversität und Inklusion mögliche Folgen und Verbesserungspotenziale unentdeckt bleiben. Wenn die Technologie sich ohne Diversitätsmaßnahmen entwickelt, ist es für einige Menschen in der Gesellschaft nicht nutzbar und sie werden exkludiert, was in gravierenden Folgen für unsere Gesellschaft und die Unternehmenskultur münden kann. Vor dem Hintergrund aktueller Technologieneuerungen ist beispielsweise die große Frage, ob KI sich hinsichtlich Diversität und Inklusion adäquat entwickelt und wenn nicht, wie man diese adressieren könnte. Die Diversity-Folgenabschätzung bietet hierbei einen bedeutsamen Ansatz, indem schon im Entwicklungsprozess über eine Implementierung vielseitiger Perspektiven und der Berücksichtigung von Diversitätsdimensionen die Folgen der Entwicklung berücksichtigt werden. Die Diversity Folgenabschätzung gliedert sich in drei Phasen, welche in der folgenden Abbildung dargestellt sind (vgl. Abbildung 1).

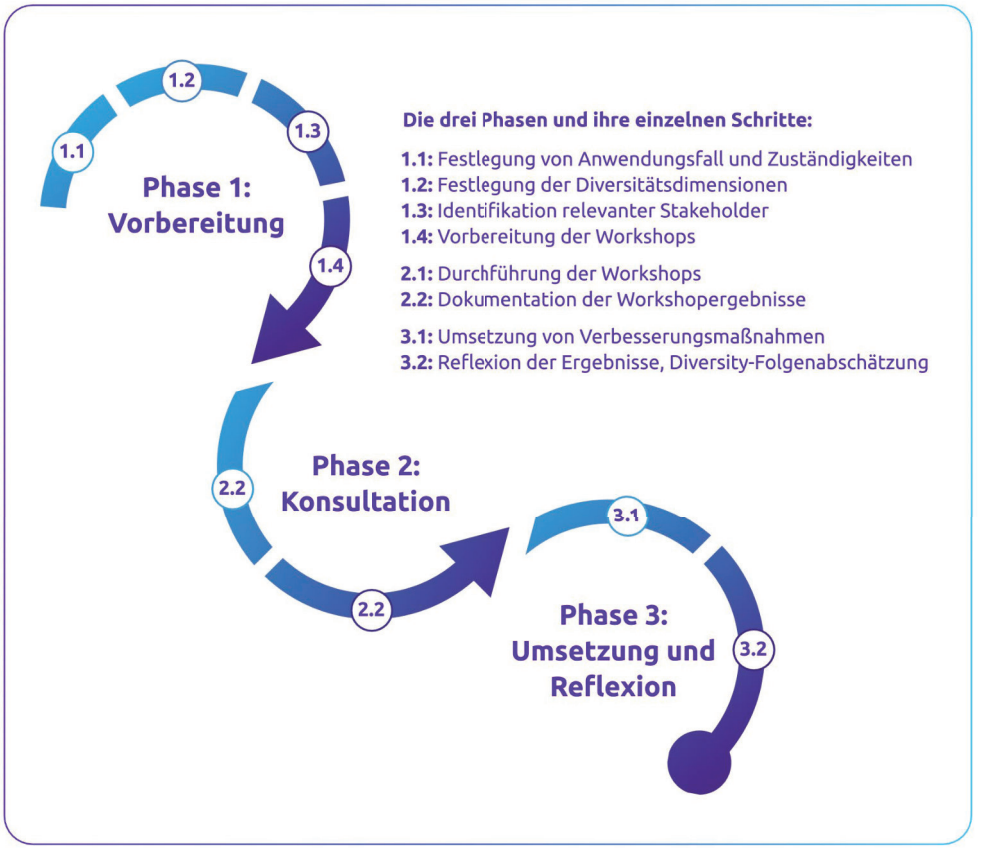


ABBILDUNG 1: PHASEN DER DIVERSITY-FOLGENABSCHÄTZUNG (QUELLE: CDR-INITIATIVE 2024: 12)

4.2 Die Diversity-Folgenabschätzung als Instrument zur Auflösung des Collingridge-Dilemmas

Wie schon erwähnt, ist das Collingridge-Dilemma ein Problem der Zeit und Technologieentwicklung, wobei es unterschiedliche Ansätze gibt, das Dilemma zu adressieren und aufzulösen. Wir möchten in diesem Kapitel zeigen, inwiefern der Diversity-Folgenabschätzung dies mit Bezug auf die Ansätze von Tigard sowie von Kudina und Verbeek gelingt.

Die Diversity-Folgenabschätzung versucht über den frühzeitigen Einbezug relevanter Stakeholder vielfältige Perspektiven in die Entwicklung einzubeziehen und damit ethische Konsequenzen früh abschätzbar zu machen. Sie werden bereits im Entwicklungsprozess über die verschiedenen Blickwinkel und ethischen Ansprüche implementiert, sodass die Technologie über den konkreten Einbezug von Diversitätsdimensionen zukünftige Folgen in der Entwicklung abzuschätzen versucht. Damit folgt die Diversity-Folgenabschätzung dem Ansatz Tigards, indem versucht wird, eine Brücke zwischen den beiden Zeitpunkten des Dilemmas zu schlagen und ethische Überlegungen in die Entwicklung einzubetten (vgl. Tigard 2022). Mit diesem Ansatz stellt sich allerdings erneut die Frage nach der Auflösung der ethischen Version des Dilemmas, d.h. jenes Dilemma, welches berücksichtigt, dass auch ethische Normen, welche implementiert werden, nicht jene sein werden, welche im späteren unsere Wertvorstellungen abbilden (vgl. Kudina/Verbeek 2019). Die Diversity-Folgenabschätzung scheint dies schwerlich vollständig auflösen zu können. Mit Blick auf den interaktiven Ansatz von Kudina und Verbeek, welcher die Wechselwirkungen von Mensch und Technologie in den Blick nimmt, lässt sich ein genauerer Blick wagen (vgl. ebd.).

Die Diversity-Folgenabschätzung der CDR-Initiative ist flexibel aufgebaut, was sich anhand vieler Feedbackprozesse und individuellen Ansätzen für das jeweilige Unternehmen und Produkt zeigt und womit es keinen einzigartigen Weg für die Förderung von Diversität gibt. Über die stetigen Reflexionsphasen und erneute Test-Phasen und Rücksprachen mit Stakeholdern kann die Diversity-Folgenabschätzung eine gewisse Dynamik adressieren und Veränderungen in Wertvorstellungen oder neue Diversitätsdimensionen einbeziehen. So können potenzielle Risiken identifiziert und notwendige Anpassungen vorgenommen werden, bevor die Technologie umfassend in die Unternehmensstrukturen integriert ist. Dennoch wird die Diversity-Folgenabschätzung in diesem Aspekt keine vollständige Auflösung des ethischen Dilemmas erreichen, da sie die Dynamik zwischen Technologie und Mensch und die daraus folgenden Veränderungen in Wert- und Inklusionsvorstellungen noch zu wenig antizipiert. Doch auch in Bezug auf die ursprüngliche Version des Collingridge-Dilemmas, d.h. dass die ethischen Konsequenzen der Technologie heute noch nicht regulierbar sind, da sie sich erst morgen zeigen, dann jedoch schwer reguliert werden können, bestehen Unzulänglichkeiten bezüglich dessen, ob die Diversity-Folgenabschätzung das Dilemma auflösen kann. Wir wollen diese im Folgenden darstellen.

5. *Herausforderungen bei der Auflösung des Collingridge-Dilemmas durch die Diversity-Folgenabschätzung*

5.1 Machtstrukturen und Hierarchien

Wir möchten in diesem Kapitel die Diversity-Folgenabschätzung, nicht ausschließlich als unternehmensinternes Vorhaben verstehen, sondern auch als ein Instrument, welches im Allgemeinen auf Technologieentwicklungen anwendbar ist. Diese umfassen damit auch auf jene, welche gesellschafts- und konsumentenorientiert sind, wie etwa die marktorientierte Entwicklung von KI-Systemen. Wir halten dies für relevant, da die Diversity-Folgenabschätzung eine gute Grundlage bietet, um ethische Überlegungen und Folgenabschätzungen schon in den Entwicklungsprozess zu integrieren und damit als Lösungsansatz für das auftretende Collingridge-Dilemma betrachtet werden kann. Die Diversity-Folgenabschätzung versucht über einen frühen Einbezug von diversen Perspektiven in den Entwicklungsprozess ethische Folgen zu erkennen und zu berücksichtigen, sodass zukünftige Folgen für die Gesellschaft und das Unternehmen abschätzbar werden und reguliert werden können. Die erfolgreiche Auflösung des Collingridge-Dilemmas durch diesen Ansatz ist jedoch an weitere Faktoren geknüpft, welchen die Diversity-Folgenabschätzung noch nicht in vollem Umfang begegnen kann und damit weiterhin Risiken bestehen Folgen in einem frühen Stadium nicht regulieren zu können, da sie in ihrer Tragweite nicht erkannt wurden.

Dies ist insbesondere in Bezug auf die Berücksichtigung unternehmensinterner Machtstrukturen der Fall. Macht und Hierarchien im Unternehmen können ein Problem für die effiziente Durchführung der Diversity-Folgenabschätzung darstellen, da in hierarchisch organisierten Unternehmen Entscheidungsträger oft großen Einfluss darauf haben, welche Risiken und ethischen Folgen überhaupt thematisiert werden. Dies kann zur Folge haben, dass bestimmte Folgen oder Betroffenheiten nicht erkannt und nicht reguliert werden, wodurch sich kostspielige und gravierende ethische Folgen bei der Implementierung der Entwicklungen ergeben können. Das Collingridge-Dilemma wird dementsprechend nicht gelöst.

In genauerer Betrachtung der von der CDR-Initiative vorgestellten Diversity-Folgenabschätzung zeigt sich dies insbesondere in der ersten Phase der Identifikation der Stakeholder und der zweiten Phase der Durchführung des Workshops, wo Machtgefälle und Hierarchien im Unternehmen ein Hemmnis der erfolgreichen Umsetzung darstellen können. In der Diversity-

Folgenabschätzung, wie sie die Initiative vorschlägt wird insbesondere auf eine verantwortungsvolle Durchführung der Folgenabschätzung basierend auf einem Top-down-Ansatz, d.h. dass die Bewertung und Umsetzung von Maßnahmen hauptsächlich von führenden Akteuren und zentralen Entscheidungsträgern gesteuert werden, gesetzt, was durchaus ihre Berechtigung hat, da Führungskräfte über eine gute Unternehmenskenntnis und Erfahrung in der Risikobewertung verfügen (vgl. CDR-Initiative 2024: 15–17). Damit besteht jedoch die Gefahr, dass Stakeholder nicht einbezogen werden, weil die Hierarchien den Blickwinkel beschränken können, so dass Stakeholder nur auf bestimmten Hierarchieebenen verortet werden. Dabei haben die Umsetzenden zwar mit den vorgeschlagenen Diversitäts-Dimensionen eine Richtlinie und Orientierung, wobei gerade diese ebenfalls eine Beschränkung darstellen kann, wenn zu wenige Dimensionen seitens der Leitenden identifiziert und für relevant befunden wurden. In Anbetracht einer erfolgreichen Abschätzung aller relevanter Folgen sollte eine Einschätzung in der Identifikation der Stakeholder sich nicht allein an der Relevanzeinschätzung der mit der Diversity-Folgenabschätzung Betrauten oder entlang der schon identifizierten Dimensionen orientieren, sondern sich hierarchiefreier gestalten. Um dies zu erreichen, sollten in der Ermittlung der relevanten Stakeholder partizipative Formate, wie breite Umfragen oder offene Gesprächsformate eingesetzt werden, sodass zunächst jedem eine Partizipationschance eingeräumt wird.

Uns ist bewusst, dass auch dies zu Verzerrungen führen kann, wenn sich bestimmte Stakeholder überproportional beteiligen und andere nicht, weshalb wir eine dynamische und zweigleisige Strategie, über eine Relevanzeinschätzung auf Grundlage der Diversitätsdimensionen seitens der mit der Folgenabschätzung betrauten Personen und zudem über partizipative Prozesse unter allen Mitarbeitern vorschlagen. Dynamisch meint hier, dass beide Prozesse miteinander im Austausch und in gegenseitiger Korrektur stehen.

Die Grundidee eines gleichberechtigten Einbezugs aller Mitarbeiter ist auch in der Workshopphase zu beachten. Hier ist vor allem bedeutsam, dass es sich um ein Kommunikationsformat argumentativen Charakters unter Beteiligung unterschiedlicher Perspektiven und Schwerpunktsetzungen, wie etwa stärker mitarbeiterorientierten Forderungen oder finanziellen Betrachtungen handelt. Um dies in einen ethischen Kontext zu setzen, spricht dies die Diskursethik von Jürgen Habermas an, die eine Orientierung an einer idealen Sprechsituation vorschlägt (vgl. Habermas 2022). Es geht darum, dass ein legitimer Konsens auf rationaler Diskussion und dem bestmöglichen Ausgleich der Interessen aller Beteiligten basiert, anstatt durch einseitige Machtinteressen

oder emotionale Überzeugungsstrategien geformt zu werden (vgl. Seiler 2014: 36). Eine ideale Sprechsituation beinhaltet demnach unter anderem eine gleichberechtigte und hierarchiefreie Kommunikationsstruktur, in welcher unabhängig von der Person die Güte des Arguments von Bedeutung ist (vgl. ebd.: 32–36). Dies ist bedeutsam, damit sich das Ergebnis nicht allein aus Partikularinteressen zusammensetzt, sondern in einem Konsens begründet liegt. In Bezug auf die Workshopphase bedeutet dies, dass eine Relevanz darin besteht, diese hierarchiefrei zu organisieren, sodass es nicht zur Unterdrückung einzelner Perspektiven und keiner Berücksichtigung kommt. Dies kann erreicht werden, indem nicht nur alle Teilnehmenden mit gegenseitigem Respekt und einer Gesprächsbereitschaft in den Dialog treten, sondern auch über eine neutrale Leitung des Diskurses durch eine Person, welche keine Anliegen in der Debatte vertritt, sondern sich der Wahrung einer gleichberechtigten Diskurssituation verschreibt.

5.2 Zukünftige Stakeholder

Während sich die CDR-Initiative in Bezug auf Hemmnisse für eine Diversity-Folgenabschätzung schon in Teilen mit Macht und Hierarchieaspekten beschäftigt, möchten wir auf einen weiteren sehr relevanten Aspekt aufmerksam machen, welcher bislang keine Berücksichtigung findet und die erfolgreiche Auflösung des Collingridge-Dilemmas insbesondere bei der Übertragung des Ansatzes der Diversity-Folgenabschätzung auf konsumentenorientierte Technologieentwicklung gefährdet. Dabei handelt es sich um die Berücksichtigung von Folgen auf zukünftige Generationen oder im unternehmensinternen Kontext auch auf zukünftige Mitarbeiter.

Unsere Beziehung zu zukünftigen Generationen ist dabei jedoch von Asymmetrien geprägt. Neben der Tatsache, dass keine gegenseitige Interaktion stattfinden kann, besteht zudem ein asymmetrisches Machtverhältnis, indem wir mit unseren Entscheidungen maßgeblichen Einfluss auf die Lebensbedingungen, als auch die Anzahl an zukünftigen Menschen haben (vgl. Meyer 2021). Seitens zukünftiger Generationen gibt es keine Handlungsmöglichkeit ihre Anliegen zu vertreten, da sie noch nicht existieren, aber auch, weil ihre Existenz in einem hohen Maße von uns abhängt. Dies bedeutet, dass wir in unseren Entscheidungen und Entwicklungen nicht nur einen Einfluss auf ihre Existenz oder ihre Lebensbedingungen haben, sondern auch darauf, was diese Menschen in Bezug auf ihre Identität ausmacht und in welchen gesellschaftlichen Strukturen sie leben (vgl. ebd.). Auch wenn ihre bloße Existenz in der Zukunft sie nicht in ein Verantwortungsverhältnis zu uns setzt, so besteht dennoch eine Verantwortung basierend auf der Annahme,

dass zukünftige Generationen basierend auf ihren zukünftigen Interessen in der Zukunft Rechte haben werden und wir diese Rechte über die Beeinflussung ihrer Interessen mit unseren Entscheidungen berühren (vgl. Hoerster 1991: 98–102).

Auch unsere Entscheidungen bezüglich der Entwicklung neuer Technologie berühren die Lebenswirklichkeit zukünftiger Generationen. Nicht nur über die pure Existenz der neuen Technologie, sondern, wie wir zu Beginn des Artikels argumentieren, auch darüber, dass die Entwicklung neuer Technologie die Strukturen und Prinzipien der Gesellschaft beeinflussen kann. Entwickeln wir beispielsweise KI-Systeme, welche auf vorurteilsbehafteten Daten basieren und diskriminierende Strukturen in der Gesellschaft über ihren Einsatz in Entscheidungsprozessen, wie der Jobvergabe dahingehend prägen, dass Chancengleichheit beispielsweise in Abhängigkeit vom Geschlecht definiert wird, so wird sich dies in einer heute noch nicht vollständig bestimmbaren Weise auf die zukünftige Gesellschaft auswirken und ihre Zuschreibung von Rechten bestimmen.

Wir tragen demnach in der Entwicklung eine Verantwortung für die Auswirkungen der Technologie auf zukünftige Generationen, wodurch, wie Emma Ruttkamp-Bloem anführt, zukünftige Generationen in der Technologieentwicklung als Stakeholder berücksichtigt werden sollten (vgl. Ruttkamp-Bloem 2023: 20–24). Im Kontext des Collingridge-Dilemmas ist dieser Aspekt höchst relevant, da die Auswirkungen unserer heutigen Entscheidungen über technologische Entwicklungen noch nicht genau bestimmbar sind. Dennoch besteht eine Verantwortung, diese Auswirkungen zu berücksichtigen, da ihre Auswirkungen oft schwer umkehrbar sind, sobald sich die Technologie etabliert hat und auch weil wir zu einem gewissen Grad beeinflussen, welche Form der gesellschaftlichen Prinzipien fortgeführt wird. Eine Auflösung des Dilemmas muss diese Auswirkungen deshalb über die Aufnahme zukünftiger Generationen als Stakeholder berücksichtigen. Dabei wird in einer der Einbeziehung von zukünftigen Stakeholdern auch eine verstärkt zukunftsorientierte Sichtweise implementiert, die Risiken nicht nur in Bezug auf uns einbezieht, sondern auch auf zukünftige Generationen.

Die Diversity-Folgenabschätzung bezieht zukünftige Generationen, oder in ihrem unternehmensinternen Kontext zukünftige Mitarbeiter*innen, nicht als Stakeholder mit ein, was dazu führt, dass die Folgenabschätzung bezüglich einer möglichen Auflösung des Collingridge-Dilemmas Schwächen aufweist. Wir geben zu, dass es eine Herausforderung darstellt Interessen und Folgen für zukünftige Generationen abzuschätzen, dennoch sehen wir eine Relevanz in diesem

Ansatz, da das Risiko verringert werden kann, dass sich problematische Technologien unumkehrbar etablieren. In einem praktischen Sinne ist die Berücksichtigung zukünftiger Stakeholder herausfordernd, da die Perspektiven zukünftiger Stakeholder, welche in der Diversity-Folgenabschätzung frühzeitig implementiert werden sollen, damit sich ethische Auswirkungen einer Technologie abschätzen lassen, noch nicht direkt bestimmbar sind. Dennoch kann wahrscheinlich allein das Bewusstsein über zukünftige Stakeholder schon eine bedeutende Rolle in der Verbesserung der Folgenabschätzung spielen, da zukünftige Risiken nicht mehr allein in Bezug auf den Nutzen für heutige Stakeholder diskontiert werden, sondern auch auf Zukünftige, was den zukünftigen Gefahren mehr Gewicht im Entscheidungsprozess verleihen würde. Im unternehmensinternen Kontext würde dies bedeuten, dass man z. B. vor dem Hintergrund des demographischen Wandels potenzielle Perspektiven zukünftiger jungen Generationen in einer alternden Gesellschaft betrachtet und in Bezug auf Diskriminierungsrisiken für Personalstrategien evaluiert. Die Diversity-Folgenabschätzung sollte demnach an die Verantwortung gegenüber zukünftigen Generationen appellieren und zukünftige Generationen als festen Stakeholder ausweisen, sodass diese Perspektive Berücksichtigung findet.

Im Kontext eines Unternehmens können im Einbezug von Interessen zukünftiger Generationen jedoch auch Anreizprobleme entstehen, da die Kosten für zukünftige Generationen nicht mehr die eigenen sind. Die Diversity-Folgenabschätzung, wird diese Anreizprobleme nicht vollständig lösen können, da sie auf der Grundlage operiert die Relevanz in der Berücksichtigung ethischer Überlegungen in der Folgenabschätzung, neben dem Verweis auf die unternehmerische Verantwortung, an unternehmenseigene Anreize, wie ihr Profitieren von einer diversen Unternehmenskultur, zu knüpfen. Während die anderen Punkte, welche für die Relevanz einer Berücksichtigung ethischer Überlegungen sprechen, sich neben der unternehmerischen Verantwortung in der Gesellschaft auch an ein Eigeninteresse knüpfen lassen, so ist letzteres in Bezug auf zukünftige Generationen schwerer zu argumentieren, was die Auflösung des Collingridge-Dilemmas gefährdet. In diesem Punkt ist es bezüglich der Auflösung des Collingridge-Dilemmas schlussendlich wahrscheinlich effektiver übergeordnete und institutionelle Regulierungen zu implementieren, um die Berücksichtigung zukünftiger Stakeholder sicherzustellen, was nicht bedeutet, dass ein Nicht-Bestehen solcher Regulierungen Unternehmen von ihrer Verantwortung entbindet. Es bedeutet lediglich, dass die Auflösung des Collingridge-Dilemmas nicht vollständig in heutigen Interessen begründet liegt und damit aus wirtschaftlichen Anreizen nur so weit bestreitbar ist, als

dass Unternehmen von einer stärker zukunftsorientierten Schätzung und Strategie profitieren werden. Dies bedeutet, dass eine Auflösung des Collingridge-Dilemmas auch dadurch geprägt ist, dass es unternehmerische als auch politische Beiträge gibt, was die Relevanz von Regularien in der Technologieentwicklung statt eines Laissez-Faire-Ansatzes verstärkt.

Die Diversity-Folgenabschätzung kann im Allgemeinen als gute Grundlage zum Versuch einer Auflösung des Collingridge-Dilemmas verstanden werden, wobei sie Verbesserungspotenzial hinsichtlich ihrer Sensitivität gegenüber Macht und Hierarchiestrukturen in einem Unternehmen als auch ihrer Zukunftsorientiertheit aufweist. Insbesondere der letzte Aspekt hat jedoch auch gezeigt, dass nicht nur die Diversity-Folgenabschätzung in der Berücksichtigung der Verantwortung zukünftiger Stakeholder vor einer Herausforderung steht, sondern ebenfalls die Möglichkeit der Auflösung des Collingridge-Dilemmas betroffen ist, da Perspektiven zukünftiger Stakeholder noch nicht direkt identifizierbar sind und damit erst in der Zukunft sichtbar sein werden.

6. Ethische Verortung

Unsere Überlegungen zu Gerechtigkeit und Diversität sowie die Prinzipien der Diversity-Folgenabschätzung lassen sich auf bedeutende ethische Hintergründe zurückführen. Bezüglich des Diskurses zur Gerechtigkeit kommt der Theorie von John Rawls eine besondere Bedeutung zu, wobei ebenso Immanuel Kants Moraltheorie und die Überlegungen Jürgen Habermas' zur demokratischen Gesellschaft Implikationen für die Unternehmensethik und Technologieentwicklung mit sich bringen. Die ethischen Aspekte der Diversity-Folgenabschätzung lassen sich demnach folgendermaßen darstellen.

Schon zu Beginn dieses Artikels haben wir die Frage, warum wir uns mit der Diversity-Folgenabschätzung befassen sollten, adressiert. Die Antwort auf diese Frage haben wir vor allem auf Überlegungen zu einer fairen Gesellschaft und unsere Verantwortung diese zu fördern gestützt, welche sich unter anderem auf die Theorie der Gerechtigkeit des amerikanischen Philosophen John Rawls zurückführen lassen. Dabei wird eine gerechte Gesellschaft mit freien und gleichen Bürgern an Prinzipien der Fairness geknüpft, die die Gleichheit der Grundfreiheiten und eine gerechte Verteilung von Ressourcen und Chancen sicherstellen sollen. Diese Prinzipien stellen ein Fundament für eine gerechte Gesellschaft dar, in der alle Bürger gleiche Chancen erhalten, unabhängig von ihrer Herkunft, sozialen Stellung oder anderen Merkmalen (vgl. Wenar 2021). In

der Umsetzung geht damit eine Verantwortung einher, systemische Barrieren abzubauen, die bestimmte Gruppen benachteiligen. Dies erfordert eine aktive Förderung von Inklusion und Gerechtigkeit, indem Ressourcen und Chancen auf eine Weise verteilt werden, die die Bedürfnisse und Rechte aller berücksichtigt. Wenn wir Rawls' Theorie auf den digitalen Bereich anwenden, dann ist ein gerechtes digitales System dasjenige, welches von allen Menschen frei und gleichberechtigt genutzt werden kann und weder im Zugang noch in der Nutzung Benachteiligungen beinhaltet. Die Fairness im Kontext der Digitalisierung und insbesondere auch in der Entwicklung von KI bedeutet damit, dass alle Menschen die gleiche Möglichkeit haben, sie zu benutzen, ohne dass Vorurteile basierend auf Geschlecht, Herkunft oder digitaler Kompetenz sie exkludieren. Wie wir gesehen haben, versucht die Diversity-Folgenabschätzung genau dies, indem sie in der Technologieentwicklung unterschiedliche Bedürfnisse berücksichtigt, um spätere Benachteiligungen zu verhindern.

Ein zweiter Aspekt unseres Beitrags besteht in der Frage, wie Unternehmen ethische Überlegungen in die Entwicklung ihrer Technologie im Sinne einer Auflösung des Collingridge-Dilemmas einbinden können. Die Diversity-Folgenabschätzung stellt dabei eine bedeutende Grundlage dar. Dabei ist jedoch zu bemerken, dass allein die Umsetzung dieser noch kein Nachkommen der moralischen Verantwortung eines Unternehmens darstellt. Laut Kant wird moralische Qualität nicht durch die Handlungen selbst bestimmt, sondern durch den guten Willen, d.h. die Absicht, die hinter einer Handlung steht. Er argumentiert, dass Handlungen nicht vollkommen gut oder schlecht sein können, sondern stattdessen ein guter oder schlechter Wille besteht (vgl. Kant 1781/1990; Hübner 2014). Anders gesagt, sollte jedes moralische Gesetz eine richtige Basis haben, wobei sein moralischer Wert genauer in der Maxime, die ihm zugrunde und in ihm verwirklicht ist, liegt (vgl. ebd.: 169). Damit liegt die Verantwortung eines Unternehmens auch darin eine ernsthafte moralische Betrachtung vorzunehmen und sich selbst an moralische Prinzipien zu binden, statt lediglich auf externe Ansprüche zu reagieren, was auch von der CDR-Initiative betont wird (vgl. CDR-Initiative 2024: 8–12). Angewandt auf den Kontext der digitalen Fairness und Diversity-Folgenabschätzung bedeutet dies, dass Unternehmen Diversität und Inklusion nicht nur als Ziel anstreben sollten, sondern dies auch aus einem echten ‚guten Willen' heraus tun. Dies ist bedeutsam, damit die Diversity-Folgenabschätzung erfolgreich sein kann und gesellschaftliche Veränderungen ernsthaft angegangen werden können.

Ein dritter Aspekt betrifft die Frage, warum wir eigentlich die Stakeholder benötigen. Könnten wir nicht lediglich eine Checkliste nutzen, um die Probleme zu identifizieren? Die Antwort auf diese Frage liegt in der Bedeutung demokratischer Prozesse und der Bedeutung des Gesprächs für eine effiziente Verbesserungskultur und Förderung von Inklusion und Diversität. Jürgen Habermas argumentiert, dass nur durch den Dialog, eine nötige Kohärenz und Legitimität der Entscheidungen erreicht werden kann, um die Herausforderungen der modernen Welt zu bewältigen und eine gerechtere, integrativere Zukunft zu gestalten (Habermas 2022). Dies bedeutet, dass angesichts der Vielfalt in unserer Gesellschaft ein kontinuierlicher Dialog notwendig ist, sodass sichergestellt ist, dass die unterschiedlichen Perspektiven in Lösungen berücksichtigt werden und keine Exklusion stattfindet. Demzufolge sind die Stakeholder wichtig für die Diversity-Folgenabschätzung, weil es ohne ihre Einbindung an demokratischen Strukturen, wie der des Meinungsaustauschs und des Einbezugs von verschiedenen Perspektiven als Korrektivfunktion fehlt. Damit stützt sich die Diversity-Folgenabschätzung auf bedeutende ethische Prinzipien, was sie zu einem wichtigen Instrument für die Förderung von Inklusion und Gerechtigkeit.

7. Schlussbemerkung

Die Berücksichtigung von Diversität und Gerechtigkeit in der Entwicklung technologischer Neuerungen stellt ein bedeutsames Thema für die Gesellschaft, als auch für Unternehmen dar, weil Technologieentwicklungen einen bedeutsamen Einfluss auf unsere gesellschaftlichen und ökonomischen Strukturen und Wertvorstellungen haben können. Politische Organisationen, wie die Europäische Union reagieren auf diese Zusammenhänge, aufgrund auch negativer und Diskriminierender Wirkungen mit ethischen Regularien zum Einsatz und Entwicklung von Technologie. Doch auch einzelnen Unternehmen kommt unabhängig von politischen Regelungen eine Verantwortung in der Implementierung ethischer Standards in der Entwicklung zur Risikominimierung zu. Ihre Herausforderung besteht dabei in der Auflösung des Collingridge-Dilemmas. Wir haben in diesem Artikel die Diversity-Folgenabschätzung, als einen Lösungsansatz betrachtet, wobei sich ergab, dass sie dieses Dilemma weitestgehend adressieren kann, insbesondere deshalb, weil sie eine Basis für eine unternehmerische Beschäftigung mit den ethischen Folgen in der Technologieentwicklung bietet und damit auch im Falle einer politischen Laissez-Faire-Situation wirksam ist. Dabei besitzt sie das Potenzial auf konsumentenorientierte

Technologieentwicklungen, statt nur unternehmensinterne Anwendungen ausgeweitet zu werden. Dennoch weist sie Schwächen in Bezug auf den Einbezug zukünftiger Stakeholder und ihrer Sensitivität gegenüber Machtstrukturen auf, welche es insbesondere auch vor dem Hintergrund der Lösung der ethischen Variante des Collingridge-Dilemmas über den Einbezug eines Bewusstseins für Werteveränderungen durch Technologie und zukünftige Stakeholder anzugehen gilt. Insgesamt stellt die Diversity-Folgenabschätzung jedoch bereits einen guten Weg in der Entwicklung inklusiver Technologien dar, indem sie versucht beide Pole des Dilemmas zu verbinden, wodurch sie einen wichtigen Schritt darstellt auch in der Zukunft die ethischen Konsequenzen der Digitalisierung zu adressieren und Unternehmen ein Instrument zur Mitarbeit an einer gerechten Gesellschaft zu bieten.

Literaturverzeichnis

O'Brian, M. / Parvini, S. (2024): How Harris and Trump Differ on Artificial Intelligence Policy, APNews, URL: https://apnews.com/article/ai-artificial-intelligence-kamala-harris-trump-vance-1c94ba2ee6a1410e61d980faa51b19e4 (aufgerufen am: 06/02/2025).

Collingridge, D. (1980): The Social Control of Technology, New York: St. Martin's Press.

CDR-Initiative (Hrsg.) (2024): Diversity-Folgenabschätzung. Praxisleitfaden zur Förderung von Diversität, Inklusion und Chancengleichheit in der Digitalisierung, Berlin, URL: https://cdr-initiative.de/news/diversity-folgenabschaetzung (aufgerufen am: 17/02/2025).

Croy, M. J. (1996): Collingridge and the Control of Educational Computer Technology, in: Techné: Research in Philosophy and Technology, Jg. 1 / Nr. 3–4, 107–115.

Dakhli, A. (2021): The Impact of Corporate Social Responsibility on Firm Financial Performance: Does Audit Quality Matter?, in: Journal of Applied Accounting Research, Jg. 23 / Nr. 5, 950–976.

Demos Helsinki (2022): What Is the Collingridge Dilemma and Why Is It Important for Tech Policy?, URL: https://demoshelsinki.fi/what-is-the-collingridge-dilemma-tech-policy/ (aufgerufen am: 05/02/2025).

Europäisches Parlament (2023): KI-Gesetz: erste Regulierung der künstlichen Intelligenz, 08.06.2023, URL: https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence (aufgerufen am: 23/03/2025).

Europäische Union (2024): Verordnung (EU) 2024/1689 des Europäischen Parlaments und des Rates vom 13. Juni 2024 zur Festlegung harmonisierter Vorschriften für künstliche

Intelligenz, in: Amtsblatt der Europäischen Union, URL: https://eur-lex.europa.eu/eli/reg/2024/1689/oj?locale=de (aufgerufen am: 23/03/2025).

Friedrich-Ebert-Stiftung (2022): Jürgen Habermas: Ein neuer Strukturwandel der Öffentlichkeit und die deliberative Politik, URL: https://www.fes.de/akademie-fuer-soziale-demokratie/buch-essenz/juergen-habermas-2022-ein-neuer-strukturwandel-der-oeffentlichkeit-und-die-deliberative-politik (aufgerufen am: 04/02/2025).

Habermas, J. (2022): Reflections and Hypotheses on a Further Structural Transformation of the Political Public Sphere, in: Theory, Culture & Society, Jg. 39 / Nr. 4, 145–171.

Hoerster, N. (1991): Abtreibung im säkularen Staat. Argumente gegen den § 218, Frankfurt: Suhrkamp.

Hübner, D. (2014): Deontologie: Das richtige Handeln, in: Hübner, D. (Hrsg.): Einführung in die philosophische Ethik, Göttingen: Vandenhoeck & Ruprecht, 151–205.

Kant, I. (1781/1990): Kritik der reinen Vernunft [KrV], hg. von Raymund Schmidt, Hamburg: Felix Meiner 1990.

Kudina, O. / Verbeek, P. (2019): Ethics from Within: Google Glass, the Collingridge Dilemma, and the Mediated Value of Privacy, in: Science, Technology, and Human Values, Jg. 44 / Nr. 2, 291–314.

Lauer, D. (2021): You Cannot Have AI Ethics Without Ethics, in: AI Ethics, Jg. 1 / Nr. 9, 21–25.

Meyer, L. (2021): Intergenerational Justice, in: Zalta. E. N. (Hrsg.): The Stanford Encyclopedia of Philosophy, URL: https://plato.stanford.edu/archives/sum2021/entries/justice-intergenerational (aufgerufen am: 17/02/2025).

Page, S. E. (2007): The Difference: How the Power of Diversity Creates Better Groups, Firms, Schools, and Societies, Princeton: Princeton University Press.

Rawls, J. (1999): A Theory of Justice: Revised Edition, Cambridge: Harvard University Press.

Ruttkamp-Bloem, E. (2023): Intergenerational Justice as Driver for Responsible AI, Conference Paper, in: Pillay, A. / Jembere, E. / Gerber, A. J. (Hrsg.): Artificial Intelligence Research, SACAIR 2023, Communications in Computer and Information Science, Jg. 1976, Cham: Springer, 12–28.

Seeliger, M./ Sevignani, S. (2022): A New Structural Transformation of the Public Sphere? An Introduction, in: Theory, Culture & Society, Jg. 39 / Nr. 4, 3–16.

Seiler, C. (2014): I Diskursethik (Habermas), in: Die Diskursethik im Spannungsfeld von Systemtheorie und Differenzphilosophie, Wiesbaden: Springer VS, 31–52.

Spiegel (2025): Trump streicht Vorgaben für Kontrolle von künstlicher Intelligenz. Trotz Warnungen vor immensen Risiken, URL: https://www.spiegel.de/netzwelt/trump-streicht-

vorgaben-fuer-kontrolle-von-kuenstlicher-intelligenz-a-fbed93e3-fe70-46f3-ab1e-891d7b81714f (aufgerufen am: 08/02/2025).

Strasser, K. / Niedermayer, B. (2021): Unvoreingenommenheit von Künstliche-Intelligenz-Systemen. Die Rolle von Datenqualität und Bias für den verantwortungsvollen Einsatz von künstlicher Intelligenz, in: Altenburger, R. / Schmidpeter, R. (Hrsg.): CSR und Künstliche Intelligenz. Management-Reihe Corporate Social Responsibility, Wiesbaden: Springer Gabler, 121–135.

Tigard, D. (2022): Embedded Ethics as Preparatory Regulation of Technology: a New Solution to the Collingridge Dilemma?, in: Lawrence, D. / Morley, S. (Hrsg.): Novel Being. Regulatory Approaches for a Future of New Intelligent Life. Elgar Law, Technology and Society series, Cheltenham: Edward Elgar Publishing, 12–28.

Wenar, L. (2021): John Rawls, in: Zalta, E. N. (Hrsg.): The Stanford Encyclopedia of Philosophy, URL: https://plato.stanford.edu/archives/sum2021/entries/rawls/ (aufgerufen am: 17/02/2025).

Fairness im Kontext der Digitalisierung[1]

Was XING von Rawls und Kant lernen kann

Paula Becker und Julian Wagner

1. Einleitung

Definitionen und Verständnisse von dem, was als fair bezeichnet werden kann, werden von Menschen unterschiedlich aufgefasst. Beispielsweise charakterisieren einige anständiges Verhalten als fair, wohingegen andere Ehrlichkeit als Kriterium für Fairness betrachten. In der praktischen Anwendung wird schnell klar, dass zwar eine Vielzahl von ethischen Intuitionen besteht, diese aber offenlassen, was genau eine Handlung unfair macht. Dies wird dann am deutlichsten, wenn ethische Probleme nicht nur benannt, sondern auch aufgelöst werden sollen. Während beispielshaft Einigkeit über die moralische Intuition existiert, dass Amazons Recruiting-Tool nicht nur Männer einstellen, und der österreichische Arbeitsmarkt-Service Frauen nicht nur Teilzeitjobs vermitteln soll, gibt es keine Klarheit darüber, wie diese und andere Probleme im Detail aufgelöst werden sollten (vgl. Wilke 2018; Köver 2024). So wird die Frage nach einer allgemeingültigen Definition immer bedeutsamer, um ethischen Schwierigkeiten wirksam entgegentreten zu können Ein besonders aktuelles Beispiel dafür bietet die Ethik der Digitalisierung. Unternehmen und Konzerne spielen eine Schlüsselrolle darin, Fairness im Kontext der Digitalisierung zu leben und zu ermöglichen. Daher ist

1 Workshop auf der #CDRK24 Konferenz unter dem Titel „Fairness in KI Richtlinien: Durchsetzbar oder leeres Versprechen?“ – geleitet von Sebastian Riemann (XING – Part of New Work SE) und Leonhard Henke (CDR-Initiative), betreut von Julian Wagner (Universität Bayreuth) und Paula Becker (Universität Bayreuth). Der nachfolgende Beitrag ist im Rahmen eines Blockseminars „Ethik der Digitalisierung“ an der Universität Bayreuth entstanden und wurde durch den Workshop auf der #CDRK24 Konferenz inspiriert. Teile dieses Fachtexts wurden unter Verwendung generativer KI-Tools erstellt. Dabei wurde Chat GPT 4.0 zur Umformulierung von Textpassagen und zur Recherche genutzt. Keinerlei inhaltliche Aspekte wurden erstellt. Alle Ergebnisse wurden fachlich überprüft und bearbeitet.

es insbesondere im unternehmerischen Kontext wichtig, ein Grundverständnis von dem, was genau mit Fairness gemeint ist, zu beleuchten. Doch was genau bedeutet Fairness im Zuge der Digitalisierung und welche unternehmerischen Verantwortungen ergeben sich daraus? Um das Problem, vor dem wir stehen, greifbarer zu machen, hier ein Beispiel: Als professionelles Business-Netzwerk dient XING dazu, geschäftliche Kontakte zu pflegen und zu erstellen. Dabei verwendet XING Algorithmen, um Bewerber*innen und Betriebe miteinander zu verbinden. Jobangebote werden passenden Kandidat*innen vorgeschlagen und umgekehrt können Unternehmen Anzeigen für offene Stellen schalten. Die Fortschrittlichkeit dieser Algorithmen wird allerdings von einem Nachteil überschattet. Denn je nachdem, mit welchen Daten der Algorithmus gefüttert wurde, kann das Ergebnis des eingesetzten Programms dazu führen, dass bestimmte Gesellschaftsgruppen weniger berücksichtigt werden. Unternehmen wie XING beschäftigen sich bereits ausführlich mit der ethischen Verantwortung in Bezug der Digitalisierung und dem Einsatz von Künstlicher Intelligenz. Was in der Diskussion innerhalb der CDR-Konferenz 2024 allerdings auffiel, ist, das zwar über Fairness diskutiert wird, dabei aber keine branchenübergreifende Einigkeit darüber besteht, was wir eigentlich meinen, wenn wir über Fairness sprechen. Das ist auch der Grund, warum die einzelnen Definitionen von Fairness zwar nicht zufällig, aber doch als individuell und gewissermaßen subjektiv bezeichnet werden können. Dieser Beitrag hat daher das Ziel, der unternehmerischen Definition von Fairness von XING eine philosophische Definition gegenüberzustellen und zu untersuchen, was XING, trotz ihres Engagements in diesem Bereich, von Rawls und Kant über Fairness und deren Umsetzung in die Praxis lernen kann.

2. Definition von Fairness am Beispiel XING

2.1 Fairness bei XING

XING konzentriert sich als professionelles Business-Netzwerk mit 22.1 Millionen registrierten Nutzer*innen auf den deutschsprachigen Raum. Dabei ermöglicht es XING, Unternehmen und Privatpersonen in Kontakt zu treten und sich auf Stellenangebote zu bewerben. Warum passiert es nun, dass bestimmte Gesellschaftsgruppen wie beispielsweise Frauen oder Menschen mit dunkler Hautfarbe Unternehmen weniger häufig vorgeschlagen werden? Wie entstehen diese Ungleichheiten? Der Grund dafür liegt häufig in dem eingesetzten Algorithmus und insbesondere in den

Daten, mit denen dieser gefüttert wurde. Die Ursache hierfür findet sich in der sogenannten „Repräsentationsverzerrung", welche beispielsweise durch Stichprobenverzerrungen verursacht wird. Das bedeutet, wenn ein Algorithmus durch die verwendeten Daten gelernt hat, dass beispielsweise Frauen weniger häufig in Ingenieurberufen eingestellt werden, dann wird er mitunter aufhören, bestimmten Unternehmen Frauenprofile vorzuschlagen, obwohl diese genauso qualifiziert sind wie ihre männlichen Mitstreiter. Wie genau löst XING diese Ungleichheiten? Um zu verstehen, wie XING mit den bestehenden Problemen von „unfairen" Algorithmen umgeht, bedarf es zunächst einer Definition von dem, was XING als Fairness definiert: "Fairness bedeutet, dass KI-Systeme Entscheidungen treffen, die „keine" unberechtigten Vorurteile oder Diskriminierungen enthalten" (Reimann 2024a) Auch erklärt XING weiter:

> Fairness in KI bedeutet, Entscheidungen ohne Vorurteile oder Diskriminierung zu treffen [...]. Für uns bei XING steht im Fokus, dass Bewerberinnen und Bewerber mit unseren KI-gestützten Recruiting Tools allein auf Basis ihrer Kompetenzen und Fähigkeiten bewertet werden – unabhängig von Alter, Herkunft oder Geschlecht. Unser Ziel: eine möglichst objektive und diskriminierungsfreie Entscheidungsgrundlage. Damit fördern wir Chancengleichheit für alle bei der Jobsuche (Reimann 2024b).

An dieser Stelle ein kurzer Hinweis. Es ist nie eine künstliche Intelligenz an sich, die unfair entscheidet. Der Ursprung einer von uns als unfair wahrgenommenen Entscheidung liegt in den Daten, mit denen der Algorithmus gefüttert wurde. Das heißt, es sind Menschen, die (bewusst oder unbewusst) dafür verantwortlich sind, dass eine künstliche Intelligenz diese unfairen Vorschläge unterbreitet und beispielhaft Menschen mit dunkler Hautfarbe Unternehmen weniger häufig vorgeschlagen werden. Da eine Verzerrung der Daten nicht immer zu vermeiden ist, bleibt zu klären, wie genau XING den Anspruch auf Nichtverzerrung umsetzt. Um trotz einer Verzerrung der Daten in KI-Systemen, deren Umgang möglichst fair und transparent zu gestalten, hat sich XING nach eigenen Angaben dazu entschieden, auf die Idee von Margaret Mitchell et al. (2019) zurückzugreifen. Diese schlagen in ihrem Beitrag aus 2019 „Model Cards for Model Reporting", eine standardisierte Methode zur Dokumentation über KI-Modellen[2] vor. Dazu erklärt XING:

---

2 Als Modell werden wir einen fertigen Algorithmus bezeichnen, der mit verschiedenen Daten trainiert wurde.

> Um Fairness in unseren KI-Systemen aktiv voranzutreiben, bereiten wir die Model-Cards nach Mitchell et al. (2019) auf. Diese dokumentieren die relevanten Eigenschaften unserer Modelle, einschließlich realer und vermuteter Biases sowie Einschränkungen, und sorgen dafür, dass mit den Modellen unternehmensintern verantwortungsvoll umgegangen wird (Reimann 2024b).

2.2 Model Cards

Um XINGs Verständnis von Fairness besser greifen zu können, lohnt es sich, etwas genauer hinter das Konzept von Model Cards zu blicken. Was genau sind Model Cards und wieso helfen sie, Fairness im Unternehmen selbst und gegenüber den Nutzer*innen von XING zu leben? Die Gründerin von TrailML, Anna Spitznagel, sieht als grundlegende Idee von Model Cards, schwer begreifliche Themen leichter zugänglich zu machen. Model Cards sollen deshalb in erster Linie die Transparenz erhöhen und die Verantwortlichkeit bei der Nutzung von KI-Modellen verbessern. Das ist deshalb so wichtig, weil viele Dinge in der Entwicklung oder in der Arbeit mit KI-Modellen verloren gehen, da Informationen nicht gespeichert oder dokumentiert werden. Doch wie genau lösen Model Cards diese Probleme und was verbirgt sich hinter diesem Konzept?

Um die Transparenz in der Entwicklung und dem Umgang mit KI-Modellen zu verbessern, werden technische Details eines Modells vorgestellt und Risikoevaluationen durchgeführt. Das gelingt, indem anhand des Schemas einer Model Card genau dokumentiert wird, wie ein Modell getestet, entwickelt und eingesetzt wird (vgl. Mitchell et al. 2019). Die Motivation dieser Dokumentationsanleitung liegt darin begründet, dass maschinelle Lernmodelle häufig in sensiblen Bereichen wie Medizin, Strafrecht oder – wie in unserem Fall – Recruiting eingesetzt werden. Dabei mangelt es oftmals an Transparenz darüber, wie genau diese Modelle funktionieren. Eine Model Card setzt hier an und stellt, wie Abbildung 1 zu entnehmen ist, als strukturiertes Dokument Informationen über Zweck, Daten, Einschränkungen, empfohlene Anwendungen und Metriken eines KI-Modells zur Verfügung.

Neben den genannten Abschnitten schlagen Mitchell et al. außerdem eine Dokumentation über „Ethical Considerations" vor (vgl. Abbildung 1). In diesem Kapitel sollen Fragen nach der Nutzung von sensiblen Daten oder möglichen Risiken von der Nutzung des KI-Modells nachgegangen werden (vgl. ebd.: 225). Durch das Protokollieren von Biases und Einschränkungen von KI-Modellen ermöglicht XING damit ein besseres, unternehmensinternes Verständnis von dem, was ihre KI-

Modelle auszeichnet und welche Schwächen es im Umgang mit ihnen zu beachten gilt. Dieses Vorgehen kann als absolut fortschrittlich beschrieben werden, da Dokumentationen über die Entstehung und Nutzung von KI-Modellen nicht verpflichtend sind. XING ist sich dabei seiner ethischen Pflicht bewusst, die Transparenz ihrer Entwicklungen im eigenen Unternehmen zu fördern und zu kommunizieren (vgl. Abbildung 1).

Aufbau einer Model Card

Model Details	Basic Information about the model (e.g. Model date, Model version, licence, person or organziation developing the model, etc.)	*Evaluation Data*	Details on the dataset(s) used for the quantitative analyses in the card
Intended Use	Use cases that were envisioned during the development	*Training Data*	May not be possible to provide in pratice. When possible, this section should mirror evaluation data. If such detail is not possible, minimal allowable information should be provided here, such as details of the distribution over various factors in the training datasets
Factors	Factor could include demographic or phenotypic groups, environmental conditions, technical attributes, etc. (e.g. primary intended uses and users, out-of-scope use cases)	*Quantitative Analyses*	Unitary results, intersectional results
Metrics	Metrics should be chosen to relfect potential real-world impacts of the model (e.g. model performance measures, variation approaches)	*Ethical Considerations* *Caveats and Recommendations*	

ABBILDUNG 1: AUFBAU EINER MODEL CARD
(QUELLE: IN ANLEHNUNG AN MITCHEL ET AL. 2019: 222)

Trotz XINGs fortschrittlichem Verhalten, sich aktiv mit Fragen über Fairness in KI-Modellen zu beschäftigen und diese in ihrem Arbeitsalltag zu integrieren, scheint dem ethischen Problem von aufkommenden Ungleichheiten auf Netzwerken wie XING insbesondere zwei Herausforderung gegenüberzustehen. Zunächst ist es wichtig anzumerken, dass die von Mitchell et al. angesprochene fehlende Transparenz durch den Einsatz von Model Cards in einem wichtigen Aspekt nicht sichergestellt wird. Große, aus aber hunderttausend und mehr Parametern bestehende Modelle sind nämlich in ihrer genauen Funktionsweise für Entwickler*innen und Nutzer*innen undurchsichtig. Dies bedeutet, dass die einzelnen Entscheidungen eines derartigen Systems nicht nachvollziehbar sind. Zum Beispiel ist es nicht möglich nachzuvollziehen, welche Input-Variablen zu einem bestimmten Output geführt haben oder noch feiner aufgelöst, welche einzelnen Bausteine bzw. Parameter innerhalb des Netzwerkes ausschlaggebend für eine Entscheidung waren (vgl. von Eschenbach 2021). So kann zum Beispiel ein neuronales Netz, welches das Kreditausfallrisiko eines bestimmten Menschen bestimmt und damit für die Allokation von Krediten zuständig ist, einem Kreditantragsteller dessen Antrag abgelehnt wurde, keine Auskunft darüber geben, welcher individuelle Faktor oder Kombination von Faktoren (Input) für die Ablehnung (Output) verantwortlich war. Dies führt dazu, dass über die einzelne Entscheidung des Modells keine hundertprozentige Vorhersage getroffen werden kann.[3]

In Model Cards finden sich deshalb nur Auswertungen über das generelle durchschnittliche Entscheidungsverhalten des Systems über viele Einzelfälle. Diese beschriebene Undurchsichtigkeit hat auch eine moralische Komponente. Wenn eine mich betreffende Entscheidung von einem undurchsichtigen System getroffen wurde, welches mir seine Abwägung für meinen Einzelfall nicht darlegen kann, wie soll ich dann Handhabe gegenüber der Maschine haben und wie möchte der Betreiber die moralische Verantwortung für die besagte Entscheidung übernehmen? Die von uns fortlaufend als „Transparenz-Challenge" gekennzeichnete Herausforderung ist also insbesondere deshalb problematisch, da Nutzer*innen von XING kaum nachvollziehen können, wie genau ein Algorithmus funktioniert und warum einem Unternehmen nun bspw. ausgerechnet dieser

[3] Das hat auch als Konsequenz, dass die Folgen von KI-Modellen nicht direkt abzuschätzen sind. Siehe zu diesem Problem den Beitrag „Technologieentwicklung und Gerechtigkeit im Zeitalter der Digitalisierung. Die Diversity-Folgenabschätzung als Instrument zur Auflösung des Collingridge-Dilemmas" von Emily Breuer und Olivia Hankins (vgl. Breuer/Hankins 2025).

Bewerber und keine Bewerberin vorgeschlagen wird (vgl. Venkatasubramanian 2020). Diese Undurchsichtigkeit auf Seiten der Nutzer*innen hat eine erhebliche moralische Komponente, auf welche wir, gerüstet mit philosophischer Theorie, später eingehen werden.

Die „Transparenz-Challenge" ist eng verknüpft mit einer weiteren Problematik: Netzwerke wie XING liefern, trotz fairen Algorithmen, unfaire Ergebnisse. Wie genau lässt sich dieser Umstand erklären? Es ist empirisch belegt, dass zwar dem Problem von unfairen Algorithmen (also Verzerrungen in den Daten) entgegengewirkt werden kann, dies löst jedoch nicht den bestehenden Bias im Endergebnis (Outcome) auf. Was genau bedeutet das? Yulia Evsyukova et al. haben in ihrer Studie „LinkedOut? A Field Experiment on Discrimination in Job Network Formation" aus 2024 gezeigt, dass Menschen mit dunkler Hautfarbe in beruflichen Netzwerken diskriminiert werden. Es konnte gezeigt werden, dass die Wahrscheinlichkeit, dass Verbindungsanfragen von Profilen von Schwarzen Menschen akzeptiert werden, um 13 Prozent geringer ist als von nicht schwarzen Profilen. Der Grund dafür liegt demnach nicht in verzerrten Daten, sondern in dem diskriminierenden Verhalten von Individuen, die sich signifikant weniger häufig dazu entscheiden, sich mit Schwarzen Profilen zu verknüpfen.

Das Ergebnis dieser Studie deckt eine zweite grundsätzliche Ebene von ethischer Problematik auf. Denn es stellt sich die Frage nach der Verantwortung von Unternehmen wie XING, wenn das Endergebnis (Outcome), trotz fairem Algorithmus, durch eine ungleiche Behandlung von Individuen mit gleichen Qualifikationen charakterisiert wird. Diese Herausforderung wird fortlaufend als „Outcome-Challenge" bezeichnet. Inwieweit dabei diese Diskriminierung noch innerhalb der Verantwortung und des Pflichtbewusstseins des Netzwerkbetreibers liegt, ist abhängig von der Auslegung und Definition von Fairness.

An dieser Stelle scheint ein erster Hinweis in Bezug zu XINGs Definition von Fairness angebracht. Denn es lässt sich anmerken, dass das Fokussieren in XINGs Fairness Definition auf den Diskriminierungsbegriff problematisch ist. So zeigt auch schon die Entwicklung von einzelnen Anwendungen, dass es nicht möglich ist, ein Modell vollständig diskriminierungsfrei zu gestalten, wenn man das Wort im Wortsinn gebraucht, wovon ohne weitere Erklärung ausgegangen werden muss. Die Aufgabe eines Machine-Learning-Systems ist es, zu diskriminieren, in dem es einteilt, einordnet oder detektiert. Die Frage sollte also sein, welche Diskriminierung moralisch zulässig ist und welche nicht. So könnte das moralisch Verwerfliche einer Diskriminierung in der Inferenz von einer Subgruppe auf eine Einzelperson liegen. Diese Ansicht betont, dass Menschen als Individuen behandelt werden sollten, nicht nur als Teile einer Gruppe.

Das Problem dieser Argumentation besteht, ähnlich wie bei der Definition von Diskriminierung darin, dass moralische Verwerflichkeit zu pauschal zugeschrieben wird. Streng genommen wäre es beispielsweise eine unzulässige Diskriminierung, wenn Bewerber*innen aufgrund eines renommierten Hochschulabschlusses als besonders produktiv im Arbeitsalltag eingestuft werden (vgl. Binns 2018). Dies basiert auf der allgemeinen Annahme, dass das Individuum die Eigenschaften der gesamten Gruppe teilt und wäre aber auch dann diskriminierend, wenn diese Annahme durch empirische Daten gestützt und im konkreten Fall zutreffend wäre. Die Nutzung des Begriffes Diskriminierung erscheint uns deshalb sowohl im Wortsinn wie er von XING benutzt wird, als auch in einer etwas genaueren Ausarbeitung in Hinsicht unsere Challenges als nicht zielführend.

Nachdem nun einige bestehende Herausforderungen für XING beleuchtet wurden, stellt sich die Frage, wie mit diesen umgegangen werden sollte. Wie sollen wir von hier weiterdenken? Unser Vorschlag ist es, sich zwei stark verschränkten philosophischen Positionen zu widmen, die uns das Nachdenken über Fairness erleichtern und uns ermöglichen, klare Ableitungen und Handlungsempfehlungen auszusprechen. Dazu entwickeln wir in den kommenden Kapiteln spezifische und philosophisch gedachte Lösungen für die „Transparenz- sowie die Outcome-Challenge".

3. Challenges

3.1 Die Transparenz-Challenge – eine kantianische Perspektive

Als Transparenz-Challenge wurde das Problem bezeichnet, das Entwickler*innen und insbesondere Nutzer*innen kaum nachvollziehen können, wie Modelle und Algorithmen im Detail funktionieren und warum Entscheidungen so getroffen werde, wie sie getroffen werden. In diesem Kapitel wollen wir aufdröseln, warum diese Bestandsaufnahme insbesondere gegenüber den Nutzer*innen von XING eine moralische Komponente enthält, die es nach Kant nicht zu unterschätzen gilt. Dazu schauen wir uns zwei wichtige Konzepte aus Kants Philosophie an und erklären anhand dieser die moralische Wichtigkeit, der „Transparenz-Challenge" entgegenzuwirken.

Kant erklärt bereits in der Vorrede der Grundlegung zur Metaphysik der Sitten (1785/2016), dass er nach einem moralischen Gesetz suchen will, das mit absoluter Notwendigkeit gelten soll. Das heißt, Moral soll, analog zu rechtlichen Gesetzen, einer Gebundenheit unterliegen. Und es

gibt gute Nachrichten: Er findet ein solches Gesetz. Kant erklärt, dass moralisches Handeln gelingt, indem sich der Mensch an ein ganz bestimmtes Prinzip hält: den kategorischen Imperativ. Doch wie genau hilft dieser bei den bestehenden Schwierigkeiten, Fairness in der Digitalisierung zu leben? Zunächst eine kleine Begriffsklärung: Ein Imperativ beschreibt erst einmal ganz grundsätzlich eine objektiv geltende Handlungsnorm. Das bedeutet, ein Imperativ schreibt nicht einfach eine Handlungsregel vor, sondern sagt aus, dass diese tatsächlich notwendig ist. Imperative besitzen also einen nötigenden Charakter. Diese Eigenschaft ist für uns von besonderem Interesse, denn unsere Kritik besteht ja unter anderem darin, dass es eine solche Nötigung für bestehende Definitionen und Umsetzungen von Fairness bisher gerade nicht gibt.

Wir begegnen also mit dem kategorischen Imperativ einer Gesetzmäßigkeit, die unabhängig von subjektivem Empfinden ist. Was bedeutet diese Entdeckung für das Konzept Fairness und dessen Umsetzung in der Unternehmenswelt? Und was genau beschreibt nun dieses von Kant entdeckte moralische Prinzip? Es gibt unterschiedliche Formen des kategorischen Imperativs, doch die allgemeinste Formel lautet: „[H]andle nur nach derjenigen Maxime, durch die du zugleich wollen kannst, daß sie ein allgemeines Gesetz werde“ (Kant 1785/2016: Abschnitt 421, Zeile 7 f.). Mit diesem berühmten Ausdruck meint Kant, dass der Mensch sich bei moralischen Überlegungen oder Unsicherheiten die Frage stellen soll, ob das Prinzip, nach dem er handelt, auch dann noch gewollt werden kann, wenn es universalisiert wird. Wenn also jeder Mensch unter allen Umständen nach eben diesem Prinzip handelt. Dabei unterziehen wir einem subjektiven Prinzip zu handeln (einer Maxime) einer Art Test. Kann der Mensch vernünftigerweise wollen, dass dieses Prinzip universalisiert wird, dann hat es den Test bestanden. Besteht die Maxime den Test nicht, dann gilt sie als unmoralisch und darf nicht ausgeführt werden. Das ist immer dann der Fall, wenn die Maxime in eine von zwei Arten von Widersprüchen gerät. Einmal darf die universalisierte Maxime keinen Widerspruch begrifflicher Art hervorrufen. Das bedeutet, das subjektive Prinzip zu Handeln darf in keinen logischen Widerspruch führen. Am Beispiel der Lüge wird das besonders deutlich. Denn wenn es meine Maxime ist, in einer Notsituation zu lügen und meinem Gegenüber zu versprechen, ihm das geliehene Geld zurückzahlen, ohne dies zu beabsichtigen, dann wird das Lügen an sich bedeutungslos. Das liegt daran, dass, wenn jeder diese Maxime verfolgen würde, die Lüge als solche nicht mehr funktioniert (vgl. ebd.: Abschnitt 422). Die Maxime hat den Test ebenso wenig bestanden, wenn dieselbe einen Widerspruch im Willen selbst herbeiführt. Hierzu ein weiteres Beispiel: Wenn ich mich dazu entscheide, meine Talente nicht zu entwickeln

und mich nicht darum bemühe, mich weiterzuentwickeln, dann kann dieses Prinzip zu Handeln zwar ohne logischen Widerspruch gedacht werden, aber Kant spricht davon, dass ich es "unmöglich wollen" (ebd.: Abschnitt 423, Zeile 28 f.) kann.

Um nun die Tragweite der „Transparenz-Challenge" zu verdeutlichen, können wir den Test auf eine beispielhaft gedachte Maxime von XING anwenden. Ein mögliches Handlungsprinzip könnte wie folgt aussehen: Solange der Algorithmus grundsätzlich faire Ergebnisse liefert, muss meinen Kund*innen nicht deutlich werden, wie diese zustande kommen. Nun fragt sich XING, kann ich wollen, dass meine Maxime ein allgemeines Gesetz wird – das also jeder Mensch nach diesem Prinzip handelt? Ist es vernünftig anzunehmen, dass faires Verhalten gelingt und umsetzbar wird, wenn diejenigen, die es betrifft, nicht darüber informiert werden, wie diese Ergebnisse zustande kommen? Wenn also Nutzer*innen gar nicht selbst darüber urteilen können, ob es fair ist, dass sie beispielshaft einem erhofften Unternehmen nicht angezeigt wurden? Es lässt sich vermuten, dass XING zu dem Schluss kommt, dass dies nicht vernünftigerweise gewollt werden kann. Denn ein solches Prinzip, zu handeln, würde für XING bedeuten, dass dem Unternehmen ebenfalls kein vollständiges Wissen über Entscheidungen zusteht, die es selbst betrifft. Das heißt, auch XING würde unter intransparenten Entscheidungen Anderer leiden. Die Undurchsichtigkeit auf der Seite der Kund*innen von XING besteht darin, dass nicht genau ersichtlich wird, wie bestimmte Verknüpfungen zustande gekommen sind. Eine mögliche Undurchsichtigkeit auf der Seite von den Mitarbeiter*innen von XING kann wiederum darin bestehen, dass XING mit Partnerfirmen zusammenarbeitet, die ihre Algorithmen nach anderen Prinzipien und ethischen Leitfäden konstruiert, wie XING das tut. Wenn keine Maxime von Transparenz besteht, dann leidet XING, in diesem Fall nun als Kunde oder Geschäftspartner, selbst unter seinem eigenen Prinzip zu Handeln. Wir begegnen einem Widerspruch im Willen selbst, da dies von XING nicht vernünftigerweise gewollt werden kann; die Maxime hat den Test also nicht bestanden.

Um zu verstehen, warum die „Transparenz-Challenge" ein echtes moralisches Problem darstellt, darf ein weiteres Konzept aus Kants Philosophie nicht fehlen: die Menschenwürde und damit einhergehend die sittliche Autonomie des Menschen (vgl. Willaschek et al. 2015: 2693). Als Würde wird bei Kant ein absoluter und nicht gegenrechenbarer Wert des Menschen bezeichnet. Diese Würde ist dem Menschen dabei praktisch zuzurechnen und weder dinglich konditioniert noch anderweitig hergeleitet. Sie ist dem Menschen qua seines Menschseins zuzuschreiben. Dabei basiert sie auf dem Gedanken der Autonomie. Das bedeutet, der Ursprung der Menschenwürde

liegt der menschlichen Fähigkeit zugrunde, sich selbst zum Handeln zu bestimmen. In dieser Funktion wird Autonomie mit Freiheit gleichgesetzt. Freiheit ist in diesem Sinne allerdings nicht so zu verstehen, dass der Mensch an kein Gesetz gebunden ist. Es ist vielmehr so, dass der Mensch an Gesetze gebunden ist, die er sich in gewisser Weise selbst auferlegt hat (vgl. Johnson/Cureton 2024). Nun scheint ein eingeschränktes Verständnis oder zumindest der Zugang zu Informationen von künstlicher Intelligenz dieser Freiheit zu widersprechen. Denn wenn nicht gewusst oder verstanden wird, wie bestimmte Entscheidungen oder Vorschläge zustande kommen, dann kann zum einen keine passende Maxime aufgestellt werden und zum anderen fehlt es den Nutzer*innen an der Möglichkeit, sich im kantischen Sinne so frei wie möglich zu entfalten und damit selbst auferlegte Gesetze adäquat zu verfolgen. Ein solches Gesetz könnte beispielshaft sein, sich in der Arbeitswelt für diejenigen Unternehmen zu entscheiden, die die eigenen Werten am besten widerspiegeln. Undurchsichtige Algorithmen können dieses Prinzip erschweren oder verhindern, wenn nicht deutlich wird, aus welchen Gründen bestimmte Unternehmen vorgeschlagen werden. Dies widerspricht aber der Autonomie des Menschen und damit zusammengehörig auch der Menschenwürde.

Das es aus moralischer Sicht also unzulässig ist, die „Transparenz-Challenge" einfach zu akzeptieren, wurde durch den Begriff der Menschenwürde und die Anwendung des kategorischen Imperativs deutlich. Denn, wenn jeder Mensch nach der beispielhaft genannten Maxime von XING handeln würde, würde es kaum gelingen, strukturellen Ungleichheiten und unfairen Ergebnissen auf Plattformen wie XING entgegenzuwirken. Auch betroffen ist dabei die Menschenwürde, insofern sich der Mensch nicht mehr adäquat zum Handeln bestimmen kann, wenn er die Ursprünge der ihn betreffenden Entscheidungen nicht nachvollziehen kann. Fragt man also Kant, sollten die Nutzer*innen nicht im Unklaren darüber bleiben, wie ein verwendeter Algorithmus Entscheidungen über Profilvorschläge etc. trifft.

3.2 Die Outcome-Challenge – eine rawlsianische Perspektive

Während sich Immanuel Kant mit der Frage von Fairness auf der Ebene des Individuums beschäftigt, stellt John Rawls die Frage nach einer fairen Gesellschaft. Was würde nun passieren, wenn sich Kantisches Denken nicht auf die Generalisierung (in Form einer Maxime) einer einzelnen Entscheidung bezieht, sondern auf die Organisation einer gesamten Gesellschaft? Diese

Frage versucht John Rawls in seinem Werk „Justice as Fairness“ zu beantworten. Um dies zu tun, vereint Rawls wesentliche Aspekte von Kants Philosophie mit seiner eigenen Vertragstheorie.

Für unsere Zwecke können wir Rawls kontraktualistische Basis auf die Idee zurückführen, dass die Einführung von gesellschaftlichen Institutionen sowie gewisse moralische Grundregeln in Form eines Gesellschaftsvertrages festgelegt werden können. Diese Grundregeln entstammen aus der Rationalität der dem Vertrag zustimmenden Personen. Das äußert sich darin, dass die abgeschlossenen Regeln dem Eigeninteresse der Einzelpersonen dienen. So argumentierte zum Beispiel schon John Locke, dass Menschen von Vernunft geleitet sind und natürliche Rechte auf Leben, Freiheit und Eigentum besitzen (vgl. Uzgalis 2024). Um diese Rechte besser zu schützen, treten sie in einen Vertrag ein, bei dem sie einer Regierung zustimmen, die auf ihrem Konsens beruht und deren Hauptaufgabe der Schutz dieser Rechte ist. Wie ein solcher Vertrag konstruiert werden soll, hängt maßgeblich von den Annahmen über die Welt und dem Menschen in dieser ab. Rawls innovativer Gedanke ist es nun, Menschen, die über die Ordnung der Gesellschaft entscheiden sollen, vor ein schon erwähntes Gedankenexperiment zu stellen.

Man stelle sich vor, man befände sich in einer Situation, in der man entscheiden soll, welche Prinzipien die Grundstruktur sowie die Institutionen einer Gesellschaft bestimmen sollen. Allerdings gibt es eine entscheidende Einschränkung: Man weiß nichts über seine eigene Stellung in dieser Gesellschaft. Man weiß nicht, ob man reich oder arm, gesund oder krank, gebildet oder ungebildet, Mann oder Frau sein wird. man kennt weder seine ethnische Zugehörigkeit noch seine Religion oder Talente. Diese Unwissenheit bezeichnet Rawls als „Schleier des Nichtwissens“. Hinter diesem Schleier des Nichtwissens haben Sie nun keine Idee davon, wie sich Ihre Entscheidungen auf Ihren persönlichen Einzelfall auswirken werden, sondern nur davon, wie sich dies auf alle möglichen Positionen auswirken könnte (vgl. Rawls 1999: 118–123).

Rawls nimmt nun an, dass Menschen hinter diesem Schleier des Nichtwissens den gleichen moralischen Ansprüchen genügen müssen, wie sie schon Kant formulierte. Um uns zu erinnern: Kants Konstruktion des kategorischen Imperativs stellt einen Vertrag des Menschen mit sich selbst dar, zu dessen Befolgung uns unsere Rationalität zwingt. Rawls hebt dies nun auf die Ebene des Gesellschaftsvertrages, welcher in den beschriebenen besonderen Umständen geschlossen wird. Welche Prinzipien ergeben sich nun daraus? Rawls formuliert zwei Prinzipien der Gerechtigkeit, nach welchen eine Gesellschaft geformt werden sollte:

> FIRST PRINCIPLE: Each person is to have an equal right to the most extensive total system of equal basic liberties compatible with a similar system of liberty for all.
>
> SECOND PRINCIPLE: Social and economic inequalities are to be arranged so that they are both: (a) to the greatest benefit of the least advantaged, consistent with the just savings principle, and (b) attached to offices and positions open to all under conditions of fair equality of opportunity (ebd.: 53).

Diese Prinzipien sind, wie schon erwähnt, ausgerichtet auf die Einrichtung der grundlegenden Institutionen einer Gesellschaft. Wie genau sollen wir nun mit diesen Prinzipien umgehen? Als erstes ist festzustellen, dass für Rawls der Gerechtigkeitsbegriff stark geknüpft ist an den Begriff der Fairness, welcher zentraler Gegenstand unseres Beitrages ist. Auch wenn Fairness nicht explizit in den beiden hier gezeigten Prinzipien erwähnt wird, sieht Rawls diese Prinzipien als zwingend, damit eine Institution fair sein kann. Weiter und expliziter formuliert er ein Prinzip der Fairness, welches auf der Ebene des Individuums seine Gültigkeit findet.

Um nun von der theoretischen philosophischen Betrachtung wieder zurück zu unserem Anwendungsfall zu kommen, ist eine Einordnung vonnöten. Wie kann XING von Rawls Prinzipien der Gerechtigkeit betroffen sein, wenn das Unternehmen keine klassische soziale Institution darstellt? Diese Frage hängt stark damit zusammen, wie wir den Begriff der Institution deuten. Auch wenn XING keine klassische Institution ist, so ist XING doch als Firma in relevanter Weise an der Verteilung von sozial bedeutsamen Gütern und auch Positionen (welche zwar nicht politische Ämter beinhalten, aber trotzdem soziale Stellungen darstellen) beteiligt (vgl. ebd.: 78–81). In dieser Weise ist das Handeln von XING gesamtgesellschaftlich relevant und hat somit Auswirkungen auf die Funktionsweise dieser Gesellschaft. Zurück unter dem Schleier des Nichtwissens ist uns zum Beispiel rationalerweise daran gelegen, dass Frauen dieselben Chancen haben, Netzwerke auf Plattformen aufzubauen wie Männer.

Wenn wir nun nach dieser Einordnung den Fall des diskriminierenden Outcomes in Bezug auf den Netzwerkaufbau von Schwarzen Profilen betrachten, sticht uns sofort Rawls zweites Prinzip der Gerechtigkeit ins Auge, welches sich mit sozialer und ökonomischer Ungleichheit auseinandersetzt. Für XINGs Handeln sind in diesem Fall beide Bausteine des Prinzips relevant. Im ersten Schritt lässt sich XINGs Wirtschaften so einordnen, dass sie durch die Zusammenführung

von Arbeitgeber*innen und Arbeitnehmer*innen einen relevanten Einfluss auf die Verteilung von Ämtern und Positionen haben. Damit unterliegen sie auf jeden Fall der Anforderung, Gleichheit der Bedingungen herzustellen. Im zweiten Schritt werden auch klare Konditionen für etwaige Ungleichbehandlungen gegeben: Ungleichbehandlung ist nur dann gerechtfertigt, wenn sie der Besserstellung der Schlechtgestelltesten dient. Die Betrachtung von ethischen Abwägungen aus der Sicht von Rawls bedeutet für XING den Perspektivwechsel in die Richtung, dass XING sich die Auswirkungen des Handelns auf die gesamte Gesellschaft vor Auge führen muss.

Durch Rawls lässt sich dies aber im zweiten Schritt schon konkret an der Einzelentscheidung (z. B. der Funktionsweise eines einzelnen Modells) festmachen. Es ist also im rawlsianischen Sinne nicht genug, dass das KI-System selbst fair entscheidet, wenn die Entscheidung keine Chancengleichheit gewährleistet. Mit Rawls gedacht, fällt also diesem, außerhalb der Entscheidungsmacht des Unternehmens liegenden Aspekt der Ungleichheit genauso viel moralisches Gewicht zu wie solchen, die dem eingesetzten Algorithmus zuzuschreiben sind. Inwieweit ist nun XING in unserem Beispiel verantwortlich für die Diskriminierung, welche von den Nutzer*innen ihrer Plattform ausgeht?

Rawls verdeutlicht, dass XING zwar keine klassische soziale Institution darstellt, jedoch aber den entsprechenden moralischen Grundsätzen gerecht werden muss. Damit einher geht auch, dass solche Institutionen moralische Ansprüche nicht nur vertreten sollen, sondern auch in ihrem Zuständigkeitsbereich für ihre Einhaltung verantwortlich sind. Dafür ist in Rawls Prinzipien auch explizit eine Kondition für Ungleichbehandlung vorgesehen, nämlich wenn sie den Schlechtgestelltesten dient. Hierzu soll erst einmal nur gesagt werden, dass es also zumindest eingeschränkt nicht gelten kann, dass XING bloß unbeteiligter und passiver Zuschauer bei „unfairen" Marktergebnissen ist, selbst wenn die eigene Vorgehensweise fair (z.B nichtdiskriminierend) ist.

Wie die „Outcome"- sowie die „Transparenz-Challenge" unserer Ansicht nach in relevanter Weise Auswirkungen auf das Handeln von XING haben sollte und welche praktischen Handlungsempfehlungen sich aus den gewonnen philosophischen Einblicken ableiten lassen, werden wir im Folgenden betrachten.

4. Was kann XING von Rawls und Kant lernen?

4.1 Handlungsempfehlungen

Transparenz praktizieren – nach dem Vorbild des kategorischen Imperativs

Um die von Kant geforderte Transparenz im Falle XINGs herzustellen, gilt es zwei Probleme zu lösen, welche auf die jeweiligen Formen der Intransparenz abzielen. Diese sind zum einen die Intransparenz gegenüber den Nutzer*innen und zum anderen die Intransparenz gegenüber den Entwickler*innen. Vorweg ist anzumerken, dass die Intransparenz von Modellen in Bezug auf beide Problematiken nur in Teilen gelöst werden kann, da ein vollständig erklärbares Modell noch außerhalb des technisch Möglichen liegt. Somit bleibt Kants Kritik gegenüber der Unmöglichkeit der Formulierung eines Imperativs beim Einsatz von Deep-Learning Modellen bestehen. Die einzige Ableitung aus diesem Fakt kann sein, erklärbare und interpretierbare Modelle zu verwenden, welche einer einsehbaren Logik folgen und damit deterministisch handeln.[4] Es gibt jedoch auch gute Nachrichten, denn mittlerweile existieren etablierte Möglichkeiten, einzelne Teilbereiche eines Modells zu erklären. Zwei explizite Methoden möchten wir beispielhaft hervorheben.

Es ist möglich, mit einer dem Entwicklungsprozess angeschlossenen Methode die sogenannte „Feature Importance" zu berechnen. Diese liefert Informationen darüber, welche Inputvariable welchen prozentualen Einfluss auf ein bestimmtes Endergebnis hat. So könnte zum Beispiel im Falle des schon betrachteten Modells, welches Kunden Kredite anhand eines Kreditausfallrisikos zuteilt, diejenigen Variablen, welche zu einer bestimmten Entscheidung geführt haben, aufgeschlüsselt und mit einer Gewichtung versehen werden (vgl. Karimi 2020).

Eine weitere Möglichkeit, in diesem Fall Transparenz herzustellen und die Handlungsfähigkeit zu gewährleisten, ist die Kommunikation mit den Nutzer*innen darüber, wie diese ihre Handlungsweisen optimalerweise anpassen können, um das Ergebnis bei einer weiteren Evaluation im gewünschten Sinne zu beeinflussen. Glücklicherweise gibt es für diese Art der Erklärung etablierte Methoden, namentlich die der „Counterfactual Explanations". Im Kontext der Kreditvergabe bedeutet dies beispielsweise, dass ein Kunde, dessen Antrag abgelehnt wurde, nicht nur erfährt, dass

4 Eine solche Ansicht vertritt zum Beispiel Cynthia Rudin, welche für den Verzicht auf undurchsichtige Modelle bei Hochrisiko-Anwendungen plädiert (vgl. Rudin 2019). Etwaige Performance Einbußen müssen laut Rudin dann entweder durch Forschung mitigiert oder akzeptiert werden.

sein Einkommen und seine Kreditwürdigkeit wesentliche Einflussfaktoren waren, sondern auch, welche konkreten Anpassungen zu einer Bewilligung geführt hätten. Eine kontrafaktische Erklärung könnte lauten: „Hätte Ihr monatliches Einkommen 500 Euro höher gelegen oder Ihr Kredit-Score um 20 Punkte besser abgeschnitten, wäre Ihr Antrag genehmigt worden". Dadurch wird nicht nur Transparenz geschaffen, sondern es entstehen auch klare Handlungsoptionen für den Antragsteller.

Dieses Beispiel zeigt, dass die genaue Gestaltung einer Erklärung individuell auf Einzelfälle angepasst werden muss, um die Sinnhaftigkeit der Information für die jeweiligen Nutzer*innen zu gewährleisten. Dieser Umstand und die Notwendigkeit, für bestimmte Anwendungen speziell entwickelte Lösungen einzusetzen macht es herausfordernd, Erklärbarkeit innerhalb XINGs Plattform zu integrieren. Während wir kein völlig ausgearbeitetes Konzept liefern können, wie XING in dem von uns angesprochenen Beispiel die Erklärbarkeit ihrer Systeme herstellen könnte, wollen wir zumindest eine Handlungsempfehlung für die generelle Herangehensweise an die Entwicklung und den Einsatz solcher Systeme liefern. Inspiration dabei könnte zum Beispiel das Framework für Erklärbarkeit von Markus Langer et al. liefern. Dieses stellt in das Zentrum der Entwicklung und des Einsatzes von Erklärbarkeitstechniken die betreffende Zielgruppe (vgl. Langer et al. 2021).

Wenn Techniken der erklärbaren Künstlichen Intelligenz (XAI) eingesetzt werden sollen, müssen nach Langer et al. verschiedene Aspekte beachtet werden. Zunächst ist es entscheidend, die relevanten Stakeholder zu identifizieren, da unterschiedliche Gruppen, etwa Nutzer*innen, Entwickler*innen oder Entscheidungsträger spezifische Anforderungen an die Erklärbarkeit haben. Diese Desiderata[5] müssen mit dem Ziel der Erklärung abgeglichen werden, um sicherzustellen, dass die bereitgestellten Informationen tatsächlich den jeweiligen Bedürfnissen entsprechen, sei es zur Erhöhung des Vertrauens, zur Handlungsermächtigung der Nutzer*innen, zur Einhaltung regulatorischer Vorgaben oder zur Verbesserung der Nutzbarkeit im Generellen.[6]

---

5 Das Wort Desiderat kommt aus dem Lateinischen und bedeutet so viel wie „Ersehntes". Ein Desiderat bezeichnet also ein Wunschobjekt.

6 Hier ist anzumerken, dass nicht nur ethische Aspekte, welche wir hier betonen, für den Einsatz erklärbarer KI sprechen, sondern auch Aspekte wie gesteigertes Vertrauen auf der Seite der Nutzer*innen, Effizienz in der Entwicklung sowie gesteigerte Kundenzufriedenheit durch solche Tools erreicht werden können.

Darauf aufbauend gilt es, den passenden Erklärungsansatz zu wählen, welcher auf die Bedürfnisse der bestimmten Stakeholder eingeht. Außerdem ist ein kontinuierlicher Feedback-Loop essenziell: Die eingesetzten Methoden müssen regelmäßig evaluiert, auf ihre Wirksamkeit geprüft und iterativ[7] verbessert werden, um den sich wandelnden Anforderungen der Stakeholder anzupassen. Es könnte eventuell nicht ganz vorhergesagt werden, welche Zielgruppe aus was für einer genauen Erklärung welchen Nutzen zieht. Zum Beispiel kommen für manche Anwendungen sowohl Feature Importance als auch Counterfactual Explanations in Frage und es muss dann iterativ eruiert werden, welche der beiden Methoden am besten funktioniert.

Entwickler*innen benötigen vor allem technisch aufgeschlüsselte Informationen zu den verschiedensten Spezifikationen eines Modells, um Modelle zu optimieren. Diese Erklärung kann und sollte sprachlich sowie inhaltlich an diese technische Expertise angepasst werden. Endnutzer*innen hingegen erwarten verständliche und nachvollziehbare Erklärungen für KI-Entscheidungen, die ihnen ermöglichen, daraus weitere Schritte abzuleiten oder entsprechend zu handeln. Diese Erklärungen müssen dann zum Beispiel in einer viel weniger technischen Sprache verfasst werden. Um dem gerecht zu werden, ist es wichtig, die Zielgruppe mit ihren Voraussetzungen (Wissen, Zeit-Ressourcen, Ansprüchen etc.) genau zu kennen und die Anwendung auf diese zuzuschneiden. Des Weiteren können Erklärungen auch auf Nachfrage der Nutzer*innen angepasst werden und damit beliebig ausführlich ausfallen, wenn dies gewünscht ist. Entwickler*innen hingegen benötigen detaillierte technische Informationen zur Optimierung des Empfehlungssystems, welche sich in Form, Sprache und Inhalt drastisch von denen der Nutzer*innen unterscheiden.

Ein weiterer Aspekt der Intransparenz betrifft die Intransparenz im Entwicklungsprozess. Einzelne Entscheidungen und Abwägungen werden iterativ im Entwicklungsprozess getroffen, jedoch nicht dokumentiert. Falls eine Dokumentation stattfindet, wie in kleinem Maße in den von XING genutzten Model Cards sind diese nur für einen engen Kreis an Entwickler*innen verständlich und dadurch auch nur für diese nutzbar. Model Cards dokumentieren jedoch auch nur den Endstand eines Modells und nicht den gesamten Entwicklungsprozess und machen es so nicht möglich, einzelne Entscheidungen innerhalb des Prozesses nachzuvollziehen. Eine solche Nachverfolgungsmöglichkeit ist in gleichem Maße notwendig, damit moralische Verantwortung für das Endprodukt und dessen Entscheidungen übernommen werden kann, genauso wie die

7 Iterativ beschreibt, dass der Prozess immer aufs Neue wiederholt wird.

Transparenz des Modells selbst. Entwicklungen in diese Richtung liefert zum Beispiel die Anwendungen des bereits erwähnten Münchner Startups TrailML. TrailML hat eine Anwendung entwickelt, die innerhalb der Entwicklungsumgebung wichtige Parameter sowie Hypothesen und Ergebnisse automatisch dokumentiert. So wird es Entwickler*innen einfach gemacht, den Entwicklungsprozess zu dokumentieren. Außerdem ist die Anwendung darauf ausgelegt, nach EU AI Act[8] konforme Berichterstattung zu leisten.

Gerechtigkeit gestalten – nach dem Vorbild des Differenzprinzips

Die für uns relevanteste Erkenntnis aus Rawls Überlegungen ist die Weitung des Blickes von Fairness innerhalb einzelner von XING eingesetzten Modellen, hin zu einer Betrachtung von Fairness im Gesamtergebnis. Entscheidend ist also die Chancengleichheit bei der Vermittlung von Arbeitnehmer*innen und Arbeitgeber*innen, welche sich im Endergebnis äußert. XING als Betreiber des Netzwerkes sollte nach Rawls als Institution entsprechend ihres Handlungsspielraums Verantwortung übernehmen. Im ersten Schritt lässt sich daraus ableiten, dass die Evaluation der von XING eingesetzten Algorithmen in Bezug auf den Fairness-Begriff auf die Handlungen der Nutzer*innen erweitert werden sollte. Dafür sind experimentelle Studien wie die von uns erwähnte Studie der Universität Mannheim entweder selbst durchzuführen oder durch entsprechende Wissenschaftler*innen durchführen zu lassen.

Wenn ein Missstand wie in der von uns angeführten Studie (vgl. Evsyukova 2025) auch im Netzwerk von XING besteht (wovon stark auszugehen ist), sollten Maßnahmen zur Mitigation[9] ergriffen werden. Nur wie sollten diese aussehen? Anzumerken ist hier der Handlungsspielraum, über den XING verfügt. Es ist nicht von XING zu erwarten, die in der Gesellschaft verankerten Vorurteile an der Wurzel des Problems zu beheben. Es ist jedoch innerhalb des Handlungsspielraums der Firma durchaus möglich, mit einem gewissen Maße an positiver Diskriminierung die Sichtbarkeit von Schwarz gelesenen Profile zu erhöhen, um Chancengleichheit herzustellen. Der Algorithmus würde also anhand einer anderen Variabel evaluiert werden, nämlich der gleichverteilten Rate von Netzwerkaufbauten für verschiedene Gruppen von Menschen. Solch eine Evaluation

8 Die Europäische Union hat neue Rechtsvorschriften zur künstlichen Intelligenz erlassen: den EU AI Act. Er schafft die Grundlagen für die Regulierung von KI in der EU.

9 Mitigation bedeutet „Abschwächung" oder „Milderung". Es geht um Maßnahmen, die die Ursachen oder Folgen eines Problems abschwächen.

durchzuführen, sollte den Entwicklern von XING höchstwahrscheinlich leichter fallen dadurch, dass sie besseren Zugriff auf die Daten des Netzwerkes haben. In das gesamte System wird dadurch jedoch auch erhebliche Komplexität eingeführt. Diese kommt daher, dass das Handeln der Nutzer*innen nun auch Auswirkungen auf die Rate der positiven Diskriminierung hat. Eine Anpassung im Algorithmus bedarf also eine hoch frequentierte Evaluation. Diesen Prozess könnten jedoch auch informierte Nutzer*innen erleichtern. Das wäre ein weiterer Vorteil der von uns im vorherigen Abschnitt erläuterten Maßnahmen zur Verringerung der Intransparenz.[10] Informierte Nutzer*innen können einsehen, ob eine Entscheidung über sie fair getroffen wurde und falls sie einen Missstand vermuten, diesen direkt effizient und genau an XING kommunizieren. Nur wenn das von XING eingesetzte System so angepasst und evaluiert wird, kann sichergestellt werden, dass es einer robusten Definition von Fairness, welche wir versucht haben dazustellen, gerecht wird.

5. Fazit

Trotz XINGs fortschrittlichem Verhalten und der Auseinandersetzung mit Fairness im Kontext der Digitalisierung konnte gezeigt werden, dass insbesondere zwei Herausforderungen bestehen bleiben: die „Transparenz"- sowie die „Outcome-Challenge". Durch die Auseinandersetzung mit den eng verknüpften Philosophien von Immanuel Kant und John Rawls wurde deutlich, dass diesen Challenges ein bedeutsames philosophisches Fundament zugrunde liegt. Um beiden Challenges entgegenzuwirken, wurden anschließend konkrete Handlungsempfehlungen ausgearbeitet, die als Lösungs- und praktische Verbesserungsvorschläge dienen können. Die „Transparenz-Challenge" kann so insbesondere dadurch gelöst oder verbessert werden, indem die Methode der „Feature Importance" oder der „Counterfactual Explanations" angewendet wird. Um die Challenge nach innen zu lösen, wurde außerdem auf das Unternehmen TrailML aufmerksam gemacht. Durch die Anwendung dieser Methoden kann es XING gelingen, mehr Transparenz gegenüber seinen Nutzer*innen zu ermöglichen. Als Lösungsvorschlag für die „Outcome-Challenge" wurde der Einsatz von Fairness-Parametern, welche über das System hinausgehen, motiviert. Namentlich die vergleichende Betrachtung des Netzwerkaufbaus von verschiedenen Nutzer*innengruppen. Mit diesem Wissen kann dann durch Rawls Argumentation positive Diskriminierung eingesetzt werden, um ein wirklich faires Endergebnis zu erzielen.

---

10 Einordnung: Nicht gegen alles kann und sollte positiv diskriminiert werden.

Binns, R. (2018): Fairness in Machine Learning: Lessons from Political Philosophy. Proceedings of the 1st Conference on Fairness, Accountability and Transparency, in: Proceedings of Machine Learning Research, Jg. 81, 149–159.

Breuer, E. / Hankins, O. (2025): Technologieentwicklung und Gerechtigkeit im Zeitalter der Digitalisierung. Die Diversity-Folgenabschätzung als Instrumetn zur Auflösung des Collingride-Dilemmas, in: Brink, A. (Hrsg.): Fairness im Zeitalter von KI, Baden-Baden: Nomos [im Erscheinen].

Evsyukova, Y. / Rusche, F. / Mill, W. (2025): LinkedOut? A Field Experiment on Discrimination in Job Network Formation, in: The Quarterly Journal of Economics, Jg. 140 / Nr. 1, 283–334, DOI: 10.1093/qje/qjae035.

Freeman, S. (Hrsg.) (2003): The Cambridge Companion to Rawls, Cambridge: Cambridge University Press.

Johnson, R. / Cureton A., (2024): Kant's Moral Philosophy, in: Zalta E. N. / Nodelman U. (Hrsg.): The Stanford Encyclopedia of Philosophy, URL: https://plato.stanford.edu/archives/fall2024/entries/kant-moral/ (aufgerufen am: 11/03/2025).

Kant, I. (1785/2016): Grundlegung zur Metaphysik der Sitten, Riga: J. F. Hartknoch.

Köver, C. (2024): Diskriminierung: AMS erntet Hohn mit neuem KI-Chatbot, in NETPOLITIK.ORG, URL: https://netzpolitik.org/2024/diskriminierung-ams-erntet-hohn-mit-neuem-ki-chatbot/#netzpolitik-pw (aufgerufen am: 25/03/2025).

Karimi, A. H. / Barthe, G. / Schölkopf, B. / Valera, I. (2020): A Survey of Algorithmic Recourse: Definitions, Formulations, Solutions, and Prospects, DOI: 10.48550/arXiv.2010.04050.

Langer, M. / Oster, D. / Speith, T. / Hermanns, H. / Kästner, L. / Schmidt, E. / Sesing, A. / Baum, K. (2021): What do we want from Explainable Artificial Intelligence (XAI)? – A Stakeholder Perspective on XAI and a Conceptual Model Guiding Interdisciplinary XAI Research, in: Artificial intelligence, Jg. 296, 103473.

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 the Conference on Fairness, Accountability, and Transparency (FAT* '19), New York: Association for Computing Machinery, 220–229, DOI: 10.1145/3287560.3287596.

Rawls, J. (1999): A Theory of Justice. Cambridge: Cambridge University Press

Reimann, S. (2024a): Vortrag 'Fairness in KI' vom 14.11.2024, Berlin [unveröffentlichte Quelle].

– (2024b): LinkedIn Korrespondenz vom 20.12.2024 [unveröffentlichte Quelle].

Rudin, C. (2019): Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead, in: Nature Machine Intelligence, Jg. 1 / Nr. 5, 206–215.

Spitznagel, A (2025): Hintergrundgespräch (29.01.2025) [unveröffentliche Quelle].

Uzgalis, W. (2024): John Locke, in: Zalta E. N. / Nodelman, U. (Hrsg.): The Stanford Encyclopedia of Philosophy (Winter 2024 Edition), URL: https://plato.stanford.edu/archives/win2024/entries/locke/ (aufgerufen am: 15/01/2025).

Venkatasubramanian, S. / Alfano, M. (2020): The Philosophical Basis of Algorithmic Recourse, in: Mireille H. / Castillo, C. / Celis, E. / Ruggieri, S. / Taylor, L. / Zanfir-Fortuna, G. (Hrsg.): Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, New York: Association for Computing Machinery, 284–293.

Von Eschenbach, W. J. (2021): Transparency and the Black Box Problem: Why We Do Not Trust AI, in: Philosophy & Technology, Jg. 34 / Nr. 4, 1607–1622.

Wilke, F. (2018): Künstliche Intelligenz diskriminiert (noch), in: Zeit Online, URL: https://www.zeit.de/arbeit/2018-10/bewerbungsroboter-kuenstliche-intelligenz-amazon-frauen-diskriminierung (aufgerufen am: 25/03/2025).

Willaschek, M. / Stolzenberg, J. / Mohr, G. / Bacin, S. (Hrsg.) (2015): Kant-Lexikon, Band 1, Berlin: De Gruyter.

KI für Alle?[1]

Implementierung von KI-Assistenten für Großunternehmen und KMU auf Basis von Interviews mit der Otto Group und des Modehaus Wellner

Julia Muff und Kilian Steinbeck

1. Einleitung

Die Digitalisierung verändert die Unternehmenslandschaft tiefgreifend und stellt Unternehmen aller Größen vor neue Herausforderungen und Chancen. Besonders die Entwicklung und Implementierung Künstlicher Intelligenz (KI) wird zunehmend zu einem entscheidenden Faktor für Wettbewerbsfähigkeit und Innovation. Während große Konzerne durch umfangreiche Ressourcen und oftmals eigenen Entwicklungsabteilungen als Vorreiter in der Nutzung von KI betrachtet werden können, stehen kleine und mittlere Unternehmen (KMU) vor der Frage, wie sie von dieser Technologie profitieren können, ohne die finanziellen und personellen Kapazitäten eines Großunternehmens zu besitzen.

Die Relevanz dieses Themas ist nicht zu unterschätzen: In Deutschland machen KMU 99 Prozent aller Unternehmen aus und sind ein zentraler Bestandteil der Wirtschaft (vgl. Statistisches Bundesamt 2022). Sie prägen den Arbeitsmarkt, fördern Innovationen und tragen maßgeblich zur wirtschaftlichen Stabilität des Landes bei. Dennoch stehen viele dieser Unternehmen vor der Herausforderung, die Potenziale von KI sinnvoll zu nutzen und gleichzeitig die Risiken und Investitionen in den Griff zu bekommen. Doch welche Lösung bietet langfristig die größeren Vorteile?

---

1 Workshop auf der #CDRK24 Konferenz unter dem Titel „Diversity-Folgenabschätzung: Ein Instrument zur Förderung von Diversität, Inklusion und Chancengleichheit" – geleitet von Franziska Nitsche (Otto Group), Thomas Bieler (ING Deutschland) und Prof. Dr. Christian Thorun (CDR-Initiative), betreut von Julia Muff (Universität Bayreuth) und Kilian Steinbeck (Universität Bayreuth). Der nachfolgende Beitrag ist im Rahmen eines Blockseminars „Ethik der Digitalisierung" an der Universität Bayreuth entstanden und wurde durch den Workshop auf der #CDRK24 Konferenz inspiriert. Teile dieses Fachtexts unter Verwendung generativer KI-Tools erstellt. Alle Ergebnisse wurden von uns fachlich überprüft und bearbeitet.

Welche strategischen, technologischen und ethischen Aspekte sind dabei zu berücksichtigen? Und welche Herausforderungen ergeben sich durch regulatorische Anforderungen oder Datenschutzrichtlinien?

Der Schwerpunkt dieses Beitrags liegt auf Unternehmen im Non-Food-E-Commerce-Sektor, da insbesondere in diesem Bereich die technische Integration sowie Optimierungspotenziale durch den Einsatz von KI eine maßgebliche Rolle spielen. So stärkt KI die Kundenloyalität, indem sie ein personalisiertes Einkaufserlebnis bietet und den Bestell- und Lieferprozess vereinfacht. Aktuelle Entwicklungen zeigen, dass insbesondere in der Modebranche im E-Commerce Künstliche Intelligenz als zentraler Innovationstreiber fungiert. Relevante Anwendungsfelder umfassen etwa personalisierte Produktempfehlungen, automatisierte Kundenkommunikation, beispielsweise durch Chatbots, visuelle Produktsuche, intelligente Lagerhaltung oder die datengetriebene Sortimentsplanung. Neben dem hohen Nutzen, etwa in Form von Effizienzsteigerung, Kostenreduktion und verbesserter Kundenzufriedenheit, bestehen jedoch auch Herausforderungen. Diese liegen insbesondere im Datenschutz, der Systemintegration sowie im Bedarf an qualitativ hochwertigen Daten. Die Bedeutung von KI für den Non-Food-E-Commerce wird damit sowohl aus strategischer als auch aus operativer Sicht zunehmend relevant (vgl. Lederer/Daus 2021).

Diese Arbeit analysiert anhand zweier Fallstudien, der Otto Group als international agierendem Großkonzern und dem Modehaus Wellner als mittelständischem Unternehmen, die unterschiedlichen Strategien zur Implementierung von KI. Dabei werden die Herausforderungen und Chancen der jeweiligen Unternehmensgröße beleuchtet. Während die Otto Group mit ihrer internen KI-Lösung „ogGPT“ auf maßgeschneiderte, datenschutzkonforme Anwendungen setzt, nutzt das Modehaus Wellner bestehende KI-Tools wie ChatGPT für ihr Marketing. Diese Gegenüberstellung ermöglicht eine differenzierte Analyse darüber, wie sich KI in Unternehmen verschiedener Größenordnungen effizient und verantwortungsvoll integrieren lässt. Neben technologischen und wirtschaftlichen Fragestellungen wird hier auch der Aspekt der Corporate Digital Responsibility (CDR) aufgegriffen. Unternehmen tragen eine gesellschaftliche Verantwortung, KI nicht nur effizient, sondern auch fair, transparent und ethisch vertretbar einzusetzen. Die Einhaltung von Datenschutzrichtlinien, die Vermeidung von Diskriminierung durch algorithmische Entscheidungen und die Sicherstellung einer menschenzentrierten Nutzung von KI sind dabei wesentliche Faktoren.

Ziel dieser Arbeit ist es, aufzuzeigen, welche Strategien KMU und Großunternehmen verfolgen können, um KI erfolgreich zu implementieren und langfristig wettbewerbsfähig zu bleiben.

Durch die Analyse praktischer Anwendungsfälle und aktueller Entwicklungen im Bereich der KI-Assistenzsysteme soll ein fundierter Überblick über die Möglichkeiten und Grenzen dieser Technologien gegeben werden. Letztlich soll die zentrale Frage geklärt werden: „Does size matter?" Ist KI ein Werkzeug, das sowohl KMU als auch Großunternehmen gleichermaßen nutzen können, oder bleibt sie eine Technologie, die erst durch umfangreiche Ressourcen wirklich rentabel wird?

Für diese Untersuchung beginnen wir mit den theoretischen Grundlagen, die auch einen Exkurs zur Entwicklung der KI umfassen (Kapitel 2). Anschließend präsentieren wir die zentralen Erkenntnisse aus den beiden geführten Interviews. Zuerst stellen wir die Ergebnisse aus dem Gespräch mit Anja Körber von der Otto Group vor, die als Vertreterin eines Großunternehmens Einblicke in die Implementierung und Nutzung ihres KI-Assistenten gibt (Kapitel 3). Danach folgen die Erkenntnisse aus dem Interview mit Holger Wellner vom Modehaus Wellner, das die Perspektive eines KMU beleuchtet (Kapitel 4). Darauf aufbauend analysieren wir die Chancen und Herausforderungen, die sich bei der Implementierung von KI in Unternehmen unterschiedlicher Größe ergeben (Kapitel 5). Abschließend widmen wir uns im Fazit der Frage, ob KI für Unternehmen aller Größen gleichermaßen als Wettbewerbsvorteil genutzt werden kann oder ob kleine und große Unternehmen jeweils spezifische Hürden zu bewältigen haben (Kapitel 6). Als Ausblick haben wir einen praktischen Leitfaden entwickelt, der Unternehmen jeder Größe praxisnahe Handlungsempfehlungen bietet, um die Potenziale von KI gezielt und erfolgreich zu nutzen (Kapitel 7).

2. Grundlagen von Künstlicher Intelligenz im E-Commerce-Kontext

Bevor wir mit dem praktischen Teil dieser Analyse beginnen und zwei Unternehmen unterschiedlicher Größen in ihrer KI-Nutzung vergleichen, ist es wichtig, die theoretischen Grundlagen der KI zu verstehen. Dazu beginnen wir mit einem historischen Exkurs zur Entwicklung der KI, um aufzuzeigen, wie sich diese Technologie von einer theoretischen Idee zu einem zentralen wirtschaftlichen und gesellschaftlichen Faktor entwickelt hat. Aufbauend auf diesem geschichtlichen Rückblick werden wir definieren, was unter Künstlicher Intelligenz verstanden wird, welche Formen existieren und wie sich deren Entwicklung, vor allem in der Non-Food-E-Commerce Branche, in den letzten Jahren beschleunigt hat. Dies legt das Fundament für die anschließende Untersuchung der spezifischen Herausforderungen und Chancen der KI-Implementierung in Unternehmen unterschiedlicher Größe.

Im Laufe der Zeit hat sich die KI von einer theoretischen Idee zu einer Schlüsseltechnologie entwickelt, die Wirtschaft und Gesellschaft grundlegend verändert. Der offizielle Startschuss für die moderne KI-Forschung erfolgte 1956 beim „Summer Research Project on Artificial Intelligence" am Dartmouth College in den USA, das von John McCarthy organisiert wurde. Zu den Teilnehmern gehörten unter anderem Marvin Minsky, Claude Shannon, Alan Newell und Herbert Simon, allesamt Visionäre, die bereits in diesen frühen Tagen der KI-Forschung überzeugt waren, dass sich Intelligenz auch unabhängig vom menschlichen Gehirn erschaffen ließe (vgl. Buxmann/Schmidt 2021). In den folgenden Jahrzehnten durchlief die KI mehrere Entwicklungsphasen. In den 1960er- und 1970er-Jahren sorgten ambitionierte Prognosen für hohe Erwartungen, die jedoch aufgrund begrenzter Rechenkapazitäten nicht erfüllt werden konnten. Dies führte zur Phase des sogenannten „KI-Winters" (vgl. Manhart 2017). In den 1980er-Jahren erlebte die KI mit regelbasierten Expertensystemen wie MYCIN, das zur Diagnoseunterstützung bei Infektionskrankheiten entwickelt wurde (vgl. Shortliffe et al. 1975), einen erneuten Aufschwung. Allerdings erwiesen sich diese Systeme langfristig als zu starr und wenig flexibel.

Die 1990er Jahre brachten mit der Verbreitung der Agententechnologie und Fortschritten in der Robotik bedeutende Neuerungen. Ein besonders prägendes Ereignis war der Schachwettkampf von 1997, in dem IBMs (International Business Machines Corporation) Supercomputer „Deep Blue" den damaligen Weltmeister Garri Kasparov besiegte. Obwohl dieser Sieg nicht als Triumph „echter" Intelligenz gewertet werden kann, da Deep Blue weder über menschenähnliches Verständnis noch kreatives Denken verfügte, zeigte er eindrucksvoll die Leistungsfähigkeit hochoptimierter Algorithmen und enormer Rechenkapazitäten. Der Schachcomputer setzte ausschließlich auf seine immense Rechenleistung und eine Vielzahl vorab programmierter Spielzüge, um optimale Entscheidungen zu treffen (vgl. Standford 2012). Mit diesem historischen Überblick wird klar, dass die KI-Entwicklung von verschiedenen Paradigmenwechseln geprägt war. Eine einheitliche Definition von Künstlicher Intelligenz ist deswegen schwer zu finden. Grundsätzlich beschreibt KI ein Teilgebiet der Informatik, das sich mit der Entwicklung intelligenter Agenten befasst. Das sind Systeme, die selbstständig Probleme lösen können (vgl. Franklin/Graesser 1996; Carbonell et al. 1983).

Heute unterscheidet man zwischen zwei Hauptarten der KI. Der schwachen KI, auch „Weak AI" oder „Narrow AI" genannt, und der starken KI, die als „Strong AI" bezeichnet wird. Während die starke KI menschliche Denkprozesse imitieren soll und Aspekte wie Bewusstsein oder Empathie anstrebt (vgl. Goertzel/Pennachin 2007; Searle 1980), bleibt sie bislang theoretischer

Natur. Die schwache KI hingegen ist bereits heute in Softwarelösungen integriert. Ihr Ziel ist es nicht, menschliches Denken nachzuahmen, sondern spezifische Problemstellungen effizient zu lösen (vgl. Goertzel/Pennachin 2007). Da eine funktionierende starke KI bislang nicht existiert und ihre Entwicklung auf unbestimmte Zeit in der Zukunft liegt, konzentriert sich diese Arbeit auf die schwache KI (vgl. Schnurr 2025). Diese ist bereits in verschiedene Softwarelösungen integriert und darauf ausgelegt, spezifische Problemstellungen effizient zu lösen. In den letzten fünf Jahren hat sich die KI-Entwicklung rasant beschleunigt, insbesondere im Bereich der „Large Language Models" (LLMs). Diese Modelle gehören zu der Kategorie des „Natural Language Processing" (NLP) und ermöglichen es der KI auf natürliche Weise mit Menschen zu kommunizieren. Ihre Stärken liegen in der Verarbeitung und Generierung von Texten, wobei sie selbstlernende Mechanismen nutzen, um Kontext und Bedeutung zu erfassen. LLMs zeichnen sich durch ihre skalierbare Architektur und den Einsatz enormer Mengen an Trainingsdaten aus. Durch diese Kombination erreichen sie ein Sprachverständnis und eine Ausdruckskraft, die mit der menschlichen Kommunikation vergleichbar sind. Aufbauend auf LLMs entstehen immer häufiger leistungsfähigere KI-gestützte Sprachwerkzeuge, die Unternehmen dabei helfen, Prozesse zu optimieren und Produktivität zu steigern. Diese Modelle haben durch Fortschritte im maschinellen Lernen und die Verfügbarkeit großer Datenmengen erhebliche Leistungssprünge gemacht (vgl. Teubner et al. 2023).

Ein bedeutender Meilenstein in der Entwicklung künstlicher Intelligenz war die Einführung von ChatGPT Ende 2022, das die Interaktion zwischen Menschen und Maschinen grundlegend veränderte. ChatGPT gehört zur Familie der GPT-Modelle („Generated Pre-trained Transformer"), die auf vortrainierten neuronalen Netzwerken basieren. Diese Modelle generieren Texte auf der Grundlage umfangreicher Trainingsdaten, wobei insbesondere GPT-3 als erstes Modell in die Kategorie der Large Language Models (LLMs) eingeordnet wurde. Grundsätzlich lassen sich zwei Arten von KI-Modellen unterscheiden: proprietäre („Black Box") und Open-Source-Modelle. Während Open-Source-Modelle den Quellcode und die zugrunde liegenden Trainingsdaten offenlegen, bleiben diese Informationen bei proprietären Modellen wie ChatGPT weitgehend undurchsichtig. Dies bedeutet, dass weder die verwendeten Trainingsdaten noch die genauen Mechanismen der Entscheidungsfindung vollständig nachvollziehbar sind, was die Interpretierbarkeit und Objektivität der generierten Inhalte beeinflussen kann (vgl. Heiser 2024).

Trotz dieser Herausforderung finden KI-Modelle wie ChatGPT zunehmend Anwendung in verschiedenen Branchen und sind aus dem beruflichen Alltag vieler Menschen nicht mehr wegzudenken. Eine aktuelle Fraunhofer-Studie zeigt, dass LLMs verstärkt in Unternehmensprozesse integriert werden, um Effizienz zu steigern und Automatisierungspotenziale auszuschöpfen (vgl. Dukino et al. 2019). Auch die Deloitte-Studie ‚Beschleunigung der KI-Transformation' belegt, dass deutsche Unternehmen zunehmend auf generative KI setzen, was ein Zeichen für das wachsende Vertrauen in diese Technologien darstellt (vgl. Bringmann et al. 2025).

Gerade im E-Commerce gehört der Einsatz von KI mittlerweile zum Standard. Die Branche gilts als Vorreiter („First Mover") in der Nutzung neuer Technologien. Laut der aktuellen Studie „The State of GenAI x Commerce", durchgeführt von Future Commerce, zeigt die schnelle Verbreitung von KI in diesem Bereich. Bereits im Jahr 2023 setzten über 90 Prozent der befragten Unternehmen KI-gestützte Assistenten ein. Zwar basiert die Studie auf der Befragung von 300 Führungskräften und Entscheidungsträger*innen aus US-amerikanischen Unternehmen im Bereich des E-Commerce, dennoch lassen sich die Ergebnisse auch auf den deutschen Markt im Non-Food-Handel übertragen. Der E-Commerce-Sektor ist international stark vernetzt, Trends und technologische Entwicklungen wie der Einsatz von generativer KI verbreiten sich besonders schnell (vgl. Future Commerce 2023).

Vor diesem Hintergrund widmet sich die vorliegende Arbeit der Frage, wie LLMs den Handelssektor, speziell die Non-Food-Branche im E-Commerce, transformieren. Indem sowohl KMU als auch Großunternehmen betrachtet werden, lassen sich branchenspezifische Entwicklungen differenziert beleuchten sowie spezifische Herausforderungen und Potenziale gezielt analysieren, um praxisrelevante Erkenntnisse zu gewinnen. Ein zentraler Bestandteil ist der Vergleich zwischen KMU und Großunternehmen im Hinblick auf die Implementierung von KI. Dabei werden bestehende Unterschiede herausgearbeitet, typische Herausforderungen identifiziert und strategische Entscheidungen betrachtet. Zudem erfolgt eine Analyse der Anforderungen und Kosten sowohl interner als auch externer KI-Assistenten, um zu bewerten, inwiefern sich der Einsatz von LLMs für Unternehmen unterschiedlicher Größenordnung rentiert. Diese theoretische Grundlage dient als Basis für die weiterführende Untersuchung bei der Implementierung von KI-Assistenten in Unternehmen.

Die Unterscheidung zwischen KMU und Großunternehmen ist essenziell für die Untersuchung der Nutzung interner und externer KI-Assistenten. Die Europäische Kommission hat im

Jahr 2003 eine offizielle Empfehlung zur Abgrenzung dieser Unternehmensgrößen veröffentlicht. Diese Definition dient nicht nur statistischen Zwecken, sondern ist auch maßgeblich für zahlreiche wirtschaftspolitische Entscheidungen, etwa im Bereich von Förderprogrammen und regulatorischen Anforderungen und deswegen hier als Abgrenzung gewählt worden. Laut dieser Definition gilt ein Unternehmen als KMU, wenn es weniger als 250 Mitarbeiter*innen beschäftigt und entweder einen Jahresumsatz von höchstens 50 Millionen Euro oder eine Jahresbilanzsumme von maximal 43 Millionen Euro aufweist. KMU spielen eine herausragende Rolle in der europäischen Wirtschaft. Mit rund 23 Millionen Unternehmen machen sie 99,8 Prozent aller Unternehmen in der EU aus und sind wesentliche Treiber für Innovation und Wettbewerbsfähigkeit. Ihre Relevanz erstreckt sich über verschiedene Sektoren hinweg, und sie profitieren insbesondere von gezielten Fördermaßnahmen und regulatorischen Erleichterungen (vgl. Europäische Kommission 2003). Im Gegensatz dazu werden Unternehmen, die mehr als 250 Mitarbeiter*innen beschäftigen und entweder einen Jahresumsatz von über 50 Millionen Euro eine Jahresbilanzsumme von mehr als 43 Millionen Euro erwirtschaften, als Großunternehmen eingestuft. Während KMU häufig durch ihre Flexibilität und Innovationskraft punkten, verfügen Großunternehmen über umfassendere Ressourcen, größere Marktanteile und eine globalere Reichweite (vgl. ebd.).

Im Rahmen dieser Untersuchung werden zwei Unternehmen als Fallbeispiele herangezogen: die Otto Group als Vertreter eines Großunternehmens und das Modehaus Wellner als KMU. Durch diese Gegenüberstellung sollen spezifische Herausforderungen und Chancen der KI-Implementierung in Unternehmen unterschiedlicher Größe analysiert werden.

3. KI im Großunternehmen: Ein Gespräch mit Anja Körber von der Otto Group

Künstliche Intelligenz ist für Unternehmen jeder Größe eine strategische Herausforderung und zugleich eine enorme Chance. Ein Beispiel für eine innovative Nutzung von KI ist die Otto Group, die mit ihrem internen KI-Assistenten „ogGPT“ neue Wege geht, um ihre Mitarbeiter*innen aktiv in den digitalen Wandel einzubinden. Anja Körber, „Head of Artitificial Intelligence (AI) & Automation“ bei der Otto Group, gibt im Interview Einblicke in die bisherigen Erfahrungen, die Herausforderungen bei der Implementierung und die Zukunftsaussichten interner KI-Assistenten.

Als international agierendes Handelsunternehmen mit Sitz in Hamburg zählt die Otto Group zu den führenden E-Commerce-Anbietern Europas. Seit ihrer Gründung im Jahr 1949 hat sie sich von einem klassischen Versandhändler zu einem breit aufgestellten Handels- und Dienstleistungskonzern entwickelt. Mit 38.456 Mitarbeiter*innen weltweit, mehr als 30 Tochtergesellschaften und einem totalen jährlichen Umsatz von 15 Milliarden Euro, setzt das Unternehmen konsequent auf digitale Innovationen und investiert gezielt in KI-gestützte Technologien (vgl. Geschäftsbericht Otto Group 2023/24).

OgGPT wurde als intern betriebene, datenschutzkonforme Lösung entwickelt, die ausschließlich innerhalb des Unternehmens läuft und die auf Sprachmodelle aus der Azure OpenAI-Umgebung zurückgreift. Azure OpenAI ist ein Cloud Service von Microsoft, der den Zugriff auf fortschrittliche KI-Modelle von OpenAI über die Azure-Plattform ermöglicht. Mit diesem Servicekönnen Unternehmen und Entwickler*innen leistungsstarke Modelle wie GPT, dessen Funktionsweise bereits in den theoretischen Grundlagen erläutert wurde, Whisper, welches dem Transkribieren dient, und DALLE, für Bildgenerierungen, nutzen (vgl. Microsoft 2025). Durch den internen Betrieb wird sichergestellt, dass die geltenden DSGVO- und Datenschutzrichtlinien eingehalten werden, während gleichzeitig leistungsstarke KI-Funktionalitäten zur Verfügung stehen. Wie Anja Körber betont (2025), ist das System barrierefrei zugänglich und kann von allen Mitarbeiter*innen genutzt werden, unabhängig von deren Rolle oder technischem Vorwissen.

Ein besonderer Fokus liegt darauf, eine vertrauenswürdige und sichere Umgebung zu schaffen, in der KI als unterstützendes Werkzeugverstanden wird. Es bietet damit eine sichere Umgebung, in der KI-Anwendungen genutzt werden können, ohne dass Unternehmensdaten für das Training externer Modelle verwendet werden könnten. Die Otto Group ist darüber hinaus Mitglied der CDR-Initiative, die sich für einen verantwortungsvollen und ethischen Umgang mit digitalen Technologien einsetzt. Die Einhaltung der Prinzipien war bei der Implementierung von ogGPT von zentraler Bedeutung und bleibt auch weiterhin ein wesentlicher Bestandteil der fortlaufenden Entwicklung (vgl. ebd.). Die Einführung des Assistenten erfolgte über einen MVP-Ansatz (Minimum Viable Product), bei dem ein schlankes, funktionales Produkt schnell bereitgestellt und kontinuierlich optimiert wurde. Genauer gesagt, ist ein MVP die früheste Version eines Produkts, die mit minimalem Aufwand entwickelt wird, um schnell am Markt getestet zu werden bzw. im Fall der Otto Group, schnell live geschalten werden kann. Ziel ist es, mit möglichst wenigen Funktionen einen realen Mehrwert für die Nutzer*innen zu bieten und durch Experimente

und Feedback die Wachstumschancen zu überprüfen. Somit können Unsicherheiten leicht reduziert und Ressourcen effizienter eingesetzt werden (vgl. Rancic/Moogk 2012).

Die erste Version von ogGPT wurde im September 2023 als eine der ersten internen KI-Lösungen eines deutschen Unternehmens live geschaltet. Von Beginn an wurde das System in kurzen circa dreiwöchigen Entwicklungszyklen weiter verbessert, um Nutzerfeedback schnell zu berücksichtigen (vgl. Körber 2025). Die Herausforderung lag weniger in der technischen Umsetzung als in der Akzeptanz durch die Belegschaft. Körber beschreibt, dass es entscheidend war, die Mitarbeiter*innen für die Nutzung zu gewinnen. Um dies zu gewährleisten, bietet die Otto Group ein breites Schulungsspektrum an – von einführenden Videos und Quick-Start-Sessions für einen schnellen Einstieg bis hin zu vertiefenden Advanced-Workshops für erfahrene Nutzer*innen. Regelmäßige Schulungen, praktische Anwendungsbeispiele und eine offene Community für Fragen haben dazu beigetragen, dass ogGPT mittlerweile in vielen Unternehmensbereichen produktiv eingesetzt wird (vgl. ebd). Besonders wichtig ist der Aspekt der AI-Adoption, wie Körber erläutert: Das Unternehmen möchte nicht nur KI-Technologie bereitstellen, sondern auch ein grundlegendes Verständnis dafür schaffen, wie sie sinnvoll eingesetzt werden kann (vgl. ebd.). Somit bezeichnet AI-Adoption die Einführung und Integration von KI in Unternehmen, wobei der Fokus nicht nur auf der technischen Implementierung liegt, sondern auch auf der Akzeptanz und Nutzung durch Mitarbeiter*innen. Ein großes Lernfeld bei der Implementierung war es, das Bewusstsein für die realistischen Einsatzmöglichkeiten und Grenzen von KI zu schärfen. Laut Körber herrscht oft die Erwartung, dass KI-Systeme sämtliche Probleme lösen oder durchgängig fehlerfreie Ergebnisse liefern. In Wahrheit beruhen große Sprachmodelle auf Wahrscheinlichkeitsbegrenzungen und sind nicht für jede Anwendung ideal. Die Otto Group legt daher Wert darauf, den Mitarbeiter*innen ein realistisches Verständnis dieser Technologie zu vermitteln und zu zeigen, wo KI sinnvoll eingesetzt werden kann und wo klassische Softwarelösungen weiterhin die bessere Wahl sind.

Ein Beispiel für diesen pragmatischen Ansatz ist die Integration eines Code-Interpreters. Während große LLMs primär mit natürlicher Sprache arbeiten, war es für viele Mitarbeiter*innen wichtig, auch mit Excel-Daten zu interagieren. Hier kommt ein separater technologischer Baustein ins Spiel: ein Code-Interpreter, der speziell für die Analyse und Visualisierung von Tabelleninhalten optimiert ist (vgl. ebd.). Dieses System kann Code direkt ausführen, Berechnungen durchführen und Daten automatisiert auswerten, ohne dass eine manuelle Programmierung erforderlich ist. So ermöglicht der Code-Interpreter eine effiziente Verarbeitung und Darstellung von Daten, was den Mitarbeiter*innen eine präzisere und interaktivere Nutzung erleichtert.

Eine weitere entscheidende Technologie ist das Prinzip der „Retrieval-Augmented Generation" (RAG). Dieses Verfahren ermöglicht es, interne Unternehmensdaten gezielt abzurufen und sie in den KI-Prozess einzubinden, ohne dass das zugrundeliegende Sprachmodell trainiert oder modifiziert werden muss. Dadurch kann ogGPT auf aktuelle, spezifische Unternehmensinformationen zugreifen, ohne die Datenschutzrichtlinien zu verletzen oder bestehende Modelle zu überlasten. Diese Kombination verschiedener KI-Technologien zeigt, dass ein modularer Ansatz entscheidend für den Erfolg eines internen KI-Assistenten ist. Darüber hinaus wird ogGPT ganz verschieden im Berufsalltag der Mitarbeitenden der Otto Group verwendet, wobei manche diese als Wissensdatenbank, manche zur Bildgenerierung oder zum Formulieren von Mails benutzen. Mit der Verbindung zu Excel-Dateien, kann das ogGPT also sowohl in Bereichen der Kommunikation, der Verwaltung, der Recherche, des Marketings, und vielen weiteren angewendet werden.

Ein weiterer entscheidender Aspekt bei der Entwicklung von ogGPT ist die Modellagnostik (vgl. ebd.). Diese bezeichnet Techniken zur Analyse und Interpretation von Machine-Learning-Modellen, die unabhängig von der Modellklasse sind. Dadurch lassen sich verschiedene Modelle bewerten, vergleichen und optimieren, ohne an eine spezifische Architektur gebunden zu sein (vgl. Ribeiro et al. 2016). Die Otto Group setzt darauf, nicht an einen einzigen Anbieter oder von einer bestimmten Modelltechnologie abhängig zu sein. Somit können neue Entwicklungen im Bereich KI flexibel und schnell integriert werden. Hierbei können die zugrundeliegenden Sprachmodelle jederzeit ausgetauscht oder erweitert werden, um sich an die neuesten technologischen Fortschritte anzupassen. Diese Agilität ist besonders relevant, da sich der Markt für KI rasant entwickelt und sich die Leistungsfähigkeit neuer Modelle kontinuierlich verbessert. Denn dank der modellagnostischen Herangehensweise kann die Otto Group schnell auf Innovationen reagieren und sicherstellen, dass ogGPT stets optimiert wird, um den höchsten Anforderungen gerecht zu werden (vgl. Körber 2025).

Für kleine und mittelständische Unternehmen stellt sich die Frage, ob eine vergleichbare interne KI-Lösung realistisch umsetzbar ist. Dies hängt stark von den jeweiligen Anforderungen und verfügbaren Mitteln ab. Körber hebt hervor, dass die Wahl zwischen interner Entwicklung und externen Lösungen sorgfältig geprüft werden sollte, weshalb eine präzise Analyse des tatsächlichen Bedarfs unerlässlich ist. Welche Aufgaben sollen mit KI unterstützt werden? Welche Anforderungen gibt es an Datenschutz und Datensicherheit? Eine fundierte Marktanalyse kann aufzeigen, ob bestehende Lösungen die Anforderungen erfüllen oder ob eine maßgeschneiderte Entwicklung notwendig ist. Für KMU mit begrenzten IT-Ressourcen, ohne ein Machine-Learning

Team, wie es die Otto Group hat, kann es sinnvoll sein, bestehende Standardlösungen mit klaren Datenschutzrichtlinien zu nutzen, anstatt eine eigene KI-Infrastruktur aufzubauen (vgl. ebd.).

Somit zeigen die Erfahrungen der Otto Group, dass interne KI-Assistenten einen entscheidenden Beitrag leisten können, um Unternehmen auf die digitale Zukunft vorzubereiten. Ein gezielter, praxisnaher Ansatz hilft Unternehmen, das Potenzial von KI sinnvoll zu nutzen und somit zukunftssicher aufzustellen.

4. KI im KMU: Ein Gespräch mit Holger Wellner vom Modehaus Wellner

Die fortschreitende Digitalisierung verändert Geschäftsprozesse grundlegend und stellt KMU vor neue Herausforderungen und Chancen. Während KI-Technologien in Großunternehmen zunehmend Anwendung finden, bleibt die Frage, inwieweit KMU bereits von diesen Innovationen profitieren. Um eine fundierte Einschätzung über den aktuellen Stand der KI-Nutzung im Mittelstand zu erhalten, wurde ein Interview mit Holger Wellner, Geschäftsführer des Modehauses Wellner, geführt.

Das 1928 gegründete Familienunternehmen mit Hauptsitz in Hameln, Niedersachsen, betreibt Filialen in Berlin und Leipzig und beschäftigt derzeit zwischen 120 und 130 Mitarbeiter*innen. Während der Großteil der Belegschaft im stationären Verkauf tätig ist, hat sich das Unternehmen in den vergangenen Jahren zunehmend digitalisiert und setzt verstärkt auf KI-gestützte Lösungen (vgl. Wellner 2025). Eine interessante Vergleichsmöglichkeit ergibt sich zur Otto Group, die als international agierendes Handelsunternehmen ebenfalls KI-gestützte Systeme einsetzt. Beide Unternehmen vertreiben Mode sowohl stationär als auch über Online-Kanäle, wodurch sich eine Untersuchung der unterschiedlichen Implementierungsstrategien im Kontext von Unternehmensgröße und Ressourcenverfügbarkeit anbietet.

Laut Holger Wellner erfolgt der Einsatz von KI im Unternehmen regelmäßig, nahezu täglich. Anstatt eine unternehmensspezifische Lösung zu entwickeln, setzt das Modehaus auf marktgängige KI-Anwendungen, die sich als leistungsfähig und wirtschaftlich effizient erwiesen. Der Schwerpunkt der KI-Nutzung liegt im Marketingbereich. Neben der automatisierten Textgenerierung nutzt das Unternehmen KI auch für die visuelle Gestaltung von Inhalten. Dabei kommt das KI-Tool MidJourney zum Einsatz, das die Erstellung von Bildern für Online-Shops und Wer-

bekampagnen unterstützt. Über das Marketing hinaus findet KI auch in der betrieblichen Verwaltung Anwendung. Sie wird unter anderem für die Analyse und Zusammenfassung von Verträgen, den Vergleich von Angeboten zur effizienteren Entscheidungsfindung sowie zur Unterstützung kreativer Prozesse eingesetzt. Beispiele hierfür sind die Generierung von Ideen für Schaufenstergestaltung oder die Planung von Events (vgl. ebd.).

Die Implementierung der Technologie erfolgte schrittweise über einen Zeitraum von etwa einem Jahr. Holger Wellner selbst kam vor eineinhalb Jahren erstmals mit ChatGPT in Berührung und erkannte rasch dessen Potenzial für den unternehmerischen Alltag. Die Einführung verlief ohne nennenswerte Hürden, da insbesondere digitalaffine Mitarbeiter*innen die KI-Technologie schnell in ihre Arbeitsabläufe integrieren konnten. Mitarbeiter*innen im stationären Verkauf, die weniger Berührungspunkte mit digitalen Prozessen haben, sind hingegen kaum betroffen (vgl. ebd.). Ein zentrales Anliegen bei der Implementierung von KI war die Produktivitätssteigerung, nicht die Substitution menschlicher Arbeitskraft. Zur erfolgreichen Integration wurde ein Schulungsprogramm in Form eines Workshops durchgeführt, der von einem externen Beratungsunternehmen begleitet wurde. Der Einsatz von KI in Unternehmen bringt jedoch nicht nur technologische und wirtschaftliche Vorteile mit sich, sondern stellt auch rechtliche und ethische Herausforderungen, insbesondere im Bereich des Datenschutzes, dar (vgl. ebd.). Als Unternehmen mit weniger als 250 Beschäftigten unterliegt das Modehaus Wellner anderen regulatorischen Anforderungen als Großunternehmen (vgl. DSGVO 2003). Dennoch werden verschiedene Maßnahmen ergriffen, um die Einhaltung datenschutzrechtlicher Bestimmungen sicherzustellen. Dazu gehören Schulungen der Mitarbeiter*innen zum sicheren Umgang mit KI-Technologien, der Einsatz der Bezahlversion von ChatGPT (vgl. Wellner 2025), die laut Anbieter eine datenschutzkonforme Verarbeitung garantiert, sowie interne Richtlinien, die die Eingabe personenbezogener oder sensibler Daten in KI-Systeme untersagen.

Im Interview wurde zudem die Frage diskutiert, welche Unterschiede zwischen KMU und Großunternehmen in Bezug auf die Nutzung von KI bestehen. Holger Wellner hebt hier insbesondere zwei Vorteile hervor, die KMU im Vergleich zu Großkonzernen haben. Während Großunternehmen durch bürokratische Strukturen gebremst werden, können KMU flexibler agieren und neue Technologien schneller implementieren. Zudem ist die Integration neuer Technologien in KMU oft unkomplizierter, da sie mit einer weniger komplexen IT-Infrastruktur arbeiten als große Konzerne mit heterogenen Systemlandschaften (vgl. ebd.).

Auf der anderen Seite stehen KMU auch vor spezifischen Herausforderungen. Die begrenzten Ressourcen für Forschung und Entwicklung erschweren die Implementierung eigener, maßgeschneiderter KI-Systeme. Allerdings sieht Holger Wellner dies nicht als entscheidenden Nachteil, da leistungsfähige externe KI-Lösungen wie ChatGPT bereits kostengünstig verfügbar sind. Für die Zukunft sieht Holger Wellner weitere Potenziale für den KI–Einsatz im Modehaus Wellner. Denkbare Einsatzfelder sind unter anderem KI-gestützte Sortimentssteuerung, um Kundenpräferenzen besser zu antizipieren, automatisierte Absatzanalysen, um den Lagerbestand optimal zu steuern, sowie intelligente Nachbestellungen, die auf realen Verkaufsdaten basieren (vgl. ebd.). Die Fallstudie zeigt, dass Künstliche Intelligenz längst im Mittelstand angekommen ist. Das Modehaus Wellner nutzt KI gezielt zur Effizienzsteigerung im Marketing und in der Verwaltung und setzt dabei auf bereits verfügbare, kosteneffiziente Lösungen. Zusammenfassend lässt sich festhalten, dass der Einsatz von KI für KMU keine Zukunftsvision mehr ist, sondern bereits Realität. Die Ergebnisse aus dem Modehaus Wellner zeigen, dass auch kleinere Unternehmen durch den gezielten Einsatz von KI effizienter arbeiten, Prozesse verbessern und sich langfristig wettbewerbsfähig aufstellen können.

5. Vergleich: Chancen und Herausforderungen für Unternehmen

5.1 Chancen

Um die unterschiedlichen Herangehensweisen an KI in Unternehmen verschiedener Größenordnungen besser zu verstehen, wurden zwei exemplarische Vertreter des Non-Food-Handels betrachtet: das Modehaus Wellner als KMU und die Otto Group als Großunternehmen. Beide setzen KI-Technologien bereits aktiv ein, jedoch auf ganz unterschiedliche Weise. Der Vergleich dieser beiden Unternehmen ermöglicht es, die zentralen Unterschiede in der Nutzung und vor allem der Implementierung von KI herauszuarbeiten.

Großunternehmen wie die Otto Group verfolgen oft langfristige, strategische Entwicklungen und können eigene KI-Lösungen nahtlos in bestehende Strukturen integrieren. KMU wie Wellner hingegen profitieren von ihrer Flexibilität und nutzen sofort einsetzbare Tools wie ChatGPT, die sich schnell und spezifisch anpassen lassen. Diese unterschiedlichen Herangehensweisen eröffnen jeweils besondere Potenziale – von Effizienzsteigerung über Automatisierung bis hin zur Individualisierung

von Kundenansprache und Prozessen. Die gewonnenen Erkenntnisse bilden zudem die Basis für praxisnahe Handlungsempfehlungen im abschließenden Kapitel. Ziel ist es, Unternehmen jeder Größe konkrete Impulse zur erfolgreichen KI-Implementierung zu geben, damit sie die Potenziale dieser Technologie noch wirkungsvoller nutzen können.

Ein zentraler Vorteil des KI-Einsatzes liegt bei beiden Unternehmen in der Automatisierung wiederkehrender Aufgaben, die Zeit und personelle Ressourcen beanspruchen. Dabei liegt aber ein zentraler Unterschied in den Anwendungsbereichen der KI-Technologie. Das Modehaus Wellner setzt KI insbesondere für Marketing und externe Kommunikation ein. Durch automatisierte Produktbeschreibungen, SEO-optimierte Texte und personalisierte Werbeinhalte kann das Unternehmen seine digitale Präsenz effektiv ausbauen, eine größere Reichweite erzielen und dies trotz begrenzter personeller Ressourcen (vgl. Wellner 2025). Im Gegensatz dazu nutzt Otto Group KI primär zur Optimierung interner Geschäftsprozesse. Der unternehmenseigene KI-Assistent ogGPT kommt in der Analyse von Daten, der Effizienzsteigerung von Arbeitsabläufen und der internen Kommunikation zum Einsatz (vgl. Körber 2025). So verbessert sich die organisationsübergreifende Zusammenarbeit, und Entscheidungen können datenbasiert getroffen werden.

Diese differenzierte Herangehensweise zeigt deutlich, dass KI nicht nur Effizienzgewinne bringt, sondern auch individuell an die jeweiligen strategischen Ziele der Unternehmen angepasst werden kann. Während die Otto Group durch maßgeschneiderte KI-Entwicklungen langfristig ihre Marktstellung sichert, können KMU wie das Modehaus Wellner durch den gezielten Einsatz externer Tools schnell auf Innovationen reagieren und effizienter arbeiten. Interessanterweise profitieren kleinere Unternehmen hier von ihrer Agilität. Ohne komplexe bürokratische Strukturen können sie neue Technologien schneller testen und implementieren. So berichtet Holger Wellner, dass sein Unternehmen innerhalb kurzer Zeit verschiedene KI-Tools eingeführt hat, darunter ChatGPT für Textgenerierung und MidJourney für Bildgestaltung. Diese Implementierung verlief reibungslos, da digital affine Mitarbeiter*innen sich schnell in die neuen Anwendungen einarbeiten konnten (vgl. Wellner 2025). Im Gegensatz dazu verfolgt die Otto Group eine langfristige, strategische KI-Integration, die eine flexible Kombination verschiedener KI-Modelle ermöglicht und eine kontinuierliche Einbindung neuer Technologien in den Unternehmensalltag sicherstellt (vgl. Körber 2025). Dieser strukturelle Unterschied zeigt, dass Großunternehmen ihre Marktstellung langfristig durch maßgeschneiderte Entwicklungen stabilisieren, während kleinere Unternehmen durch ihre Agilität und schnelle Entscheidungsprozesse besonders innovativ sein können.

Zudem profitieren sowohl KMU als auch Großunternehmen von der Fähigkeit der KI-Integration, interne Prozesse zu optimieren und vorhandene Daten besser zu nutzen. Somit nutzt die Otto Group ihre interne KI, ogGPT, um Mitarbeiter*innen bei der Analyse von Geschäftsprozessen und der internen Kommunikation zu unterstützen. Durch die gezielte Nutzung von KI-Technologien kann das Unternehmen effizientere Abläufe schaffen und datenbasierte Entscheidungen erleichtern (vgl. ebd.). Auch das Modehaus Wellner setzt auf KI-gestützte Systeme, insbesondere im Marketing und in der Verwaltung. KI steigert auch hier die Effizienz und ermöglicht gezielte Ansprache der Kundschaft (vgl. Wellner 2025).

Neben den wirtschaftlichen Vorteilen spielt auch die ethische und transparente Nutzung von KI eine zunehmend wichtige Rolle. Unternehmen müssen nicht nur die Effizienz steigern, sondern auch sicherstellen, dass ihre KI-Systeme fair, nachvollziehbar und datenschutzkonform eingesetzt werden. Die Otto Group setzt mit ogGPT auf eine datenschutzkonforme Lösung, die sicherstellt, dass sensible Unternehmensdaten geschützt bleiben und nicht für das Training externer Modelle genutzt werden. Damit geht das Unternehmen bewusst einen Schritt in Richtung CDR, indem es KI-Technologien nutzt, ohne dabei die Kontrolle über interne Daten aufzugeben. Der Fokus liegt auf einem sicheren und transparenten Einsatz, der Mitarbeiter*innen das Potenzial der Technologie zugänglich macht, ohne dabei auf rechtliche oder ethische Risiken einzugehen (vgl. Körber 2025). Das Modehaus Wellner legt großen Wert auf eine verantwortungsvolle Integration von KI, fördert dazu jedoch eine flexible und auf Vertrauen basierende Vorgehensweise. Die Nutzung der Bezahlversion von ChatGPT, die strengere Datenschutzbestimmungen erfüllt, trägt zusätzlich zur Sicherheit bei. Um das Wissen um den sicheren Umgang mit KI zu fördern, wurde kürzlich eine Schulung eingeführt, die das Verständnis für diese Technologie stärkt. Diese Schulungen werden künftig intensiviert, um die Mitarbeiter*innen weiterhin zu unterstützen und die Integration von KI im Unternehmensalltag zu optimieren (vgl. Wellner 2025).

Die Erfahrungen der Otto Group und des Modehauses Wellner zeigen, dass Unternehmen jeder Größe Maßnahmen ergreifen können, um KI ethisch und transparent zu nutzen. Die Entwicklung von verbindlichen Leitlinien, die Schulung von Mitarbeiter*innen und die sorgfältige Auswahl datenschutzkonformer Lösungen sind zentrale Faktoren, um die Chancen der KI-Technologie zu nutzen, ohne dabei Risiken einzugehen. Letztendlich trägt ein verantwortungsvoller KI-Einsatz nicht nur zur Einhaltung regulatorischer Standards bei, sondern stärkt auch das Vertrauen von Kundschaft und Mitarbeiter*innen in die digitale Transformation. Somit zeigt der

Vergleich beider Unternehmen, dass sowohl Großunternehmen als auch KMU enorme Chancen durch den gezielten Einsatz von KI haben, indem zum Beispiel ihre Abläufe optimiert, neue Prozesse entwickelt und ihre Wettbewerbsfähigkeit gestärkt werden können.

5.2 Herausforderungen

Die Implementierung von KI-Technologien eröffnet Unternehmen zwar neue Möglichkeiten, bringt jedoch auch eine Vielzahl an Herausforderungen mit sich. Diese unterscheiden sich je nach Unternehmensgröße deutlich. Während Großunternehmen wie die Otto Group in der Regel über die nötigen Ressourcen und Infrastrukturen verfügen, stoßen sie häufig auf interne Widerstände, wie etwa bei der Akzeptanz der neuen Technologien durch Mitarbeiter*innen, stehen KMU oft vor ganz anderen Schwierigkeiten. So müssen sie etwa mit begrenztem Fachwissen, knappen Budgets oder fehlenden IT-Infrastrukturen umgehen. Die folgenden Beispiele zeigen, wie sich diese unterschiedlichen Ausgangsbedingungen konkret auf die Umsetzung von KI-Projekten auswirken. In diesem Unterkapitel werden die Herausforderungen analysiert, mit denen Unternehmen bei der Implementierung von KI–Assistenten konfrontiert sind.

Eine der zentralen Herausforderungen für sowohl das Modehaus Wellner als auch die Otto Group ist der Datenschutz. Sobald Daten in eine KI eingespeist werden, müssen klare Regeln für deren Verarbeitung und Speicherung existieren, um Datenschutzverstöße zu vermeiden. Die Otto Group begegnet dieser Herausforderung, indem sie ihren eigenen KI-Assistenten entwickelt hat, der höchsten Datenschutzstandards entspricht. In Kombination mit der Nutzung von Azure OpenAI stellt das Unternehmen sicher, dass sensible Unternehmensdaten nicht für das Training externer Modelle verwendet werden (vgl. Körber 2025). Das Modehaus Wellner hingegen setzt primär auf Eigenverantwortung der Mitarbeiter*innen, daher dürfen sie keine datenschutzrechtlich kritischen Informationen in den KI-Chat eingeben. Die Server von OpenAI erfüllen dabei die geltenden Datenschutzrichtlinien (vgl. Wellner 2025).

Eine weitere wesentliche Herausforderung für Unternehmen jeder Größe ist die Akzeptanz von KI–Technologien durch die Mitarbeiter*innen. Die technische Umsetzung selbst ist dabei meist nicht das größte Hindernis. Vielmehr geht es darum, die Belegschaft davon zu überzeugen, KI–gestützte Assistenzsysteme aktiv in ihren Arbeitsalltag zu integrieren. Die Otto Group begegnet diesem Problem mit umfassenden Schulungsangeboten, die praxisnahe Anwendungsbeispiele

vermitteln und offene Fragen der Mitarbeiter*innen klären. Ziel ist es, mögliche Vorbehalte abzubauen und den Mehrwert der Technologie verständlich zu machen (vgl. Körber 2025). Auch Holger Wellner ist sich dieser Herausforderung bewusst, sieht in seinem Unternehmen jedoch weniger Widerstand gegenüber der Einführung von KI. Im Modehaus Wellner arbeiten nur jene Mitarbeitenden mit der Technologie, die ohnehin Interesse daran haben. Insbesondere für die Beschäftigten im Verkauf, die den Großteil der Belegschaft ausmachen, besteht keine Verpflichtung, KI in ihre Arbeitsprozesse zu integrieren. Dennoch setzt auch das Modehaus Wellner auf gezielte Schulungen, um die Akzeptanz zu fördern und den Mitarbeiter*innen die Potenziale der Technologie näherzubringen (vgl. Wellner 2025).

Eine weitere wesentliche Herausforderung für KMU liegt in den begrenzten Ressourcen für die Entwicklung und Implementierung von KI-Technologien im Vergleich zu Großunternehmen. Dies wurde auch im Vergleich zwischen der Otto Group und dem Modehaus Wellner deutlich. Während Otto einen eigenen KI-Assistenten auf Basis eines bestehenden Modells entwickelt hat, entschied sich das Modehaus für eine ressourcenschonendere Lösung und nutzt ChatGPT Premium (vgl. ebd.). Dadurch kann die Anpassung an unternehmensspezifische Anforderungen für das Modehaus schwieriger sein, da es über weniger Ressourcen zur individuellen Weiterentwicklung der KI verfügt als ein Großkonzern wie Otto. Allerdings lässt sich aus dem Interview mit Holger Wellner nicht eindeutig ableiten, dass dies tatsächlich als Herausforderung wahrgenommen wird. Er betont, dass die bestehenden KI-Tools nicht nur kostengünstig, sondern auch gut nutzbar sind. Seiner Einschätzung nach stellt der Einsatz einer standardisierten KI-Lösung für sein Unternehmen daher keine signifikante Einschränkung dar.

Holger Wellner sieht die größeren Herausforderungen beim Einsatz von KI weniger in seinem eigenen Unternehmen, sondern eher aufseiten des Großunternehmens Otto. Ein wesentlicher Faktor sind die höheren regulatorischen Hürden, die dazu führen, dass Veränderungen langsamer umgesetzt werden können als in einem KMU. Dies könnte zunehmend an Bedeutung gewinnen, da sich die KI-Technologie sehr schnell weiterentwickelt und kontinuierliche Anpassungen erforderlich sind. Wenn große Unternehmen aufgrund bürokratischer Strukturen langsamer auf Neuerungen reagieren können, stellt dies eine erhebliche Herausforderung dar, die aktiv bewältigt werden muss (vgl. ebd.). Die Otto Group begegnet dieser Problematik, indem sie Prozesse in den Arbeitsalltag integriert hat, die eine schnelle Umsetzung von Änderungen ermöglichen. Ein Beispiel dafür ist die regelmäßige Aktualisierung ihres internen KI-Assistenten ogGPT, die alle drei Wochen erfolgt.

Durch diese strukturierte Vorgehensweise stellt Otto sicher, dass der Assistent stets auf dem neuesten Stand bleibt und sich an veränderte Anforderungen anpassen kann (vgl. Körber 2025). Die Integration von KI-Assistenzsystemen in bestehende IT-Strukturen stellt insbesondere für Großunternehmen eine Herausforderung dar, da deren IT-Landschaften häufig komplex und stark vernetzt sind. Die Otto Group meisterte diese Aufgabe, indem sie GPT strukturiert in die bestehenden Systeme integrierte, wodurch eine effiziente Nutzung sichergestellt wurde (vgl. ebd.). Im Gegensatz dazu war die Implementierung im Modehaus Wellner deutlich einfacher, da die IT-Strukturen dort weniger komplex sind. Während große Unternehmen oft mit heterogenen Systemen arbeiten, die eine aufwendige Anpassung neuer Technologien erfordern, sind die IT-Systeme in kleineren Unternehmen meist überschaubarer und lassen sich schneller erweitern. Dies führt zu einer deutlich geringeren technischen Hürde bei der Einführung von KI-Lösungen in KMU.

Ein zentraler Erfolgsfaktor bei der Implementierung von KI-Assistenzsystemen liegt daher darin, die vielfältigen Herausforderungen – von Datenschutz und Akzeptanz von Mitarbeiter*innen bis hin zu begrenzten Ressourcen und komplexer Systemintegration – sowohl auf technischer als auch auf organisatorischer und strategischer Ebene proaktiv anzugehen, um die jeweiligen Unternehmensgegebenheiten zu berücksichtigen.

6. Fazit

Die Analyse der KI-Implementierung in der Otto Group und im Modehaus Wellner zeigt, dass sowohl Großunternehmen als auch KMU von künstlicher Intelligenz profitieren können, jedoch mit unterschiedlichen Strategien und Herausforderungen konfrontiert sind. Während die Otto Group durch ihre umfassende IT-Infrastruktur maßgeschneiderte, strategische KI-Lösungen entwickelt, nutzt das Modehaus Wellner flexible, schnell implementierbare KI-Tools zur Optimierung des Marketings und der externen Kommunikation.

Ein zentraler Unterschied liegt in der Herangehensweise: Großunternehmen wie die Otto Group verfolgen eine langfristige Integration von KI in interne Prozesse, was zwar nachhaltige Vorteile bringt, aber aufgrund bürokratischer Strukturen oft langsamer erfolgt. KMU hingegen können von ihrer Agilität profitieren, indem sie ohne aufwendige interne Abstimmungsprozesse schnell neue Technologien einführen können. Dies ermöglicht es Unternehmen wie dem Modehaus Wellner, rasch auf Marktveränderungen zu reagieren und innovative KI-Tools effizient zu

nutzen. Zudem stellt die Integration von KI in bestehende IT-Strukturen insbesondere für Großunternehmen eine größere Herausforderung dar, während KMU von einer weniger komplexen IT-Landschaft profitieren.

Trotz dieser strukturellen Unterschiede gibt es auch gemeinsame Herausforderungen. Datenschutz und die Akzeptanz der Mitarbeiter*innen sind für beide Unternehmen zentrale Themen. Während die Otto Group mit ogGPT eine datenschutzkonforme Inhouse-Lösung entwickelt hat, setzt das Modehaus Wellner auf Sensibilisierung und Eigenverantwortung der Mitarbeiter*innen im Umgang mit KI. Dabei ist der Erfolg nicht nur eine Frage der Technologie, sondern auch der Strategie und Unternehmenskultur. Transparente Kommunikation, gezielte Schulungen und ein realistischer Blick auf die Stärken und Grenzen von KI sind entscheidend, um Mitarbeiter*innen auf diese Reise mitzunehmen. Unternehmen, die sich frühzeitig mit KI auseinandersetzen und bewusst gestalten, wie sie diese Technologie einsetzen wollen, können langfristig davon profitieren.

Die Gegenüberstellung der Otto Group und des Modehauses Wellner zeigt somit, dass sich die Frage „Does size matter?" klar beantworten lässt: Size does not matter! Folglich entscheidet die Unternehmensgröße allein nicht über den Erfolg von KI-Strategien. Denn während Großunternehmen durch maßgeschneiderte KI-Lösungen langfristig ihre Marktstellung sichern, können KMU durch ihre Flexibilität und den gezielten Einsatz externer Tools effizienter und innovativer arbeiten. Beide Unternehmen demonstrieren, dass der richtige Einsatz von KI zu Produktivitätssteigerungen, besseren Entscheidungsprozessen und neuen Wettbewerbsvorteilen führt. Entscheidend ist eine maßgeschneiderte KI-Strategie, die technologische, wirtschaftliche und ethische Aspekte gleichermaßen berücksichtigt – und das unabhängig von der Größe des Unternehmens. Die Zukunft gehört also den Unternehmen, die sich aktiv mit KI auseinandersetzen und die Chancen der Digitalisierung gezielt nutzen.

7. Ausblick: Ein Leitfaden zur erfolgreichen Implementierung von KI in Unternehmen

Die durchgeführten Interviews mit der Otto Group und dem Modehaus Wellner haben eindrücklich gezeigt, dass Unternehmen unterschiedlichster Größe bereits aktiv in die Implementierung von KI investieren. Dabei treten weder unüberwindbare Hürden noch fundamentale Schwierigkeiten auf, die den Zugang zu KI-Technologien einschränken würden. Vielmehr zeigen die Erfahrungen beider Unternehmen, dass mit einer durchdachten Strategie und einem klaren Ansatz die Potenziale von KI erfolgreich ausgeschöpft werden können. Denn die Möglichkeiten, KI in

Unternehmen einzusetzen, sind nahezu unbegrenzt. Sie sind ein zentraler Baustein der digitalen Transformation und eröffnen Organisationen jeder Größe neue Chancen. Egal, ob Großkonzern oder KMU, wer die Potenziale der KI erkennt und für sich nutzt, kann Effizienz steigern, neue Geschäftsmodelle erschließen und Wettbewerbsvorteile sichern. Doch wie gelingt der Einstieg in diese Technologie, und welche Schritte sind für eine erfolgreiche Implementierung entscheidend? Körber hat im Interview bereits einen kurzen Ausblick für die Nutzung von KI für KMU erörtert (vgl. Körber 2025), doch wie sollte die Implementierung grundsätzlich ausschauen, unabhängig von der Unternehmensgröße?

Dieses letzte Kapitel bietet somit einen Ausblick mit einem umfassenden Leitfaden zur erfolgreichen Implementierung von KI-Technologien, der aufzeigt, wie Unternehmen ihre spezifischen Stärken nutzen und individuelle Anforderungen berücksichtigen können. Durch gezielte Strategien und praxisnahe Ansätze wird veranschaulicht, wie der richtige Einsatz von KI nicht nur die Effizienz steigert, sondern auch neue Geschäftsmodelle ermöglicht und die Wettbewerbsfähigkeit stärkt. Unternehmen sind somit nicht nur gefordert, sich mit KI auseinanderzusetzen, sondern können auch aktiv von den digitalisierten Möglichkeiten profitieren, die diese transformative Technologie zu bieten hat. Um KI nachhaltig in Unternehmensprozesse zu integrieren, ist eine strukturierte Vorgehensweise entscheidend. Die Einführung sollte in vier zentrale Phasen unterteilt werden:

1. Zielsetzung und Folgenabschätzung,
2. Planung und Gestaltung,
3. Vorbereitung und Implementierung sowie
4. Evaluation und Anpassung (vgl. Stowasser et al. 2020).

Viele Unternehmen haben bereits KI-basierte Lösungen erfolgreich in ihre Prozesse integriert (vgl. Bringmann et al. 2025). Dabei zeigt sich, dass es wichtig ist, individuelle Anforderungen zu berücksichtigen und nicht einfach bestehende Modelle zu kopieren. Eine durchdachte Strategie und gezielte Implementierung sind entscheidend, um langfristig von den Vorteilen der KI zu profitieren Die zuvor genannten Phasen bilden dabei dieses methodische Vorgehen (vgl. Stowasser et al. 2020). Die Phase 1 umfasst die Entwicklung einer klaren Strategie als Basis jeder erfolgreichen Implementierung. Unternehmen müssen zunächst diese grundlegenden Fragen zu klären: Welchen Mehrwert kann KI bieten? Welche Prozesse lassen sich durch KI effizienter gestalten? Welche konkreten Anwendungsbereiche stehen im Fokus? (vgl. Körber 2025).

Entscheidend ist ein problemorientierter Ansatz, der sich nicht allein an technologischen Möglichkeiten orientiert, sondern an den spezifischen Herausforderungen des Unternehmens. Klare, gut definierte Use Cases ermöglichen erste Erfolge und wertvolle Lernprozesse. Ein schrittweises Vorgehen minimiert Risiken und schafft nachhaltige Ergebnisse (vgl. Dukino et al. 2019). Darüber hinaus ist auch die Unternehmenskultur von hoher Relevanz. Die Entwicklung von ethischen KI-Leitlinien hilft, Transparenz zu schaffen und den Einsatz der Technologie nachhaltig zu gestalten. Interne Kommunikationsmaßnahmen und kontinuierliche Feedbackmechanismen spielen eine zentrale Rolle beim Aufbau von Vertrauen in KI-Systeme (vgl. Stowasser et al. 2020). Dies zeigt sich auch am Beispiel der Otto Group, die laut Körber (2025) entsprechende Maßnahmen implementiert.

Die Phase 2 beinhaltet neben der strategischen Planung und Gestaltung auch die frühzeitige Einbindung der Unternehmenskultur und der internen Kommunikation ein entscheidender Erfolgsfaktor. Eine transparente Darstellung der Ziele, Chancen und Risiken der KI-Einführung stärkt das Vertrauen und die Akzeptanz innerhalb der Belegschaft. Um eine reibungslose Integration der Technologie zu gewährleisten, sollten mögliche Bedenken frühzeitig erkannt und aktiv adressiert werden.

Die Phase 3 erfordert einen strukturierten Implementierungsprozess. Ein. Zunächst gilt es, Geschäftsprozesse zu analysieren und jene Abläufe zu identifizieren, die sich besonders für den KI-Einsatz eignen. Der Aufbau interdisziplinärer Teams aus internen und externen Expert*innenerleichtert die Entwicklung passender Lösungen (vgl. Dukino et al. 2019). Die ersten Pilotprojekte sollten sich auf Anwendungen mit hohem Nutzen und geringer Komplexität konzentrieren, um rasche Erfolge zu ermöglichen. Die daraus gewonnenen Erkenntnisse dienen als Fundament für eine nachhaltige Skalierung und die Entwicklung neuer Geschäftsmodelle. Dieser Ansatz entspricht dem MVP-Ansatz der Otto Group: Zunächst wird ein grundlegendes Modell geschaffen, das im Laufe des Prozesses kontinuierlich optimiert wird (vgl. Körber 2025). Denn Unternehmen, die frühzeitig Change-Management-Strukturen implementieren, profitieren langfristig durch effizientere Prozesse, innovative Geschäftsmodelle und eine gestärkte Wettbewerbsfähigkeit (vgl. Stowasser et al. 2020).

Die letzte und vierte Phase, die die Evaluation und Anpassung des Systems beinhaltet, berücksichtigt dass KI kein statisches System, sondern ein kontinuierlicher Lernprozess ist. Der Erfolg von KI-Anwendungen sollte regelmäßig anhand zuvor definierter Zielsetzungen und

Nutzenversprechen bewertet werden. Sowohl Erfolge als auch Fehlschläge liefern wertvolle Erkenntnisse zur Optimierung der Strategie. Open-Source- und Standard-KI-Tools ermöglichen maßgeschneiderte Lösungen, während eine systematische Abarbeitungsreihenfolge dafür sorgt, dass sowohl einfache als auch komplexe Anwendungen sinnvoll integriert werden (vgl. Dukino et al. 2019).

Die Einführung von Künstlicher Intelligenz ist für Unternehmen jeder Größe ein bedeutender Schritt in Richtung Zukunft. Entscheidend ist eine strukturierte Herangehensweise – von der Strategieentwicklung über die gezielte Implementierung bis hin zur kontinuierlichen Optimierung. KI kann betriebliche Prozesse optimieren und neue Geschäftsmodelle ermöglichen. Unternehmen, die frühzeitig in KI investieren und eine klare Strategie verfolgen, sichern sich langfristig Wettbewerbsvorteile (vgl. Stowasser et al. 2020). Wer die Chancen der digitalen Transformation aktiv nutzt, genau wie das Modehaus Wellner und die Otto Group, kann Effizienz steigern, Innovation vorantreiben und seine Marktposition langfristig stärken. So wird Künstliche Intelligenz nicht nur zu einem technologischen Fortschritt, sondern zu einem entscheidenden Erfolgsfaktor in der modernen Wirtschaft.

Literaturverzeichnis

Bringmann, B. / Fach, P. / Becker, S. J. / Schamberger, M. / Wendland, P. (2025): Now Decides Next: Generating a New Future. Deloitte's State of Generative AI in the Enterprise Quarter Four Report – German Cut March 2025, URL: https://www.deloitte.com/de/de/Industries/technology/research/ki-studie.html (aufgerufen am: 30/03/2025).

Buxmann, P. / Schmidt, H. (2021): Grundlagen der Künstlichen Intelligenz und des Maschinellen Lernens, in: Buxmann, P. / Schmidt, H. (Hrsg.): Künstliche Intelligenz, Berlin: Springer, 3–25.

Carbonell, J. G. / Michalski, R. S. / Mitchell, T. M. (1983): An Overview of Machine Learning, in: Machine Learning, Jg. 1 / Nr. 1, 3–23.

Chaib-Draa, B. / Moulin, B. / Mandiau, R. / Millot, P. (1992): Trends in Distributed Artificial Intelligence, in: Artificial Intelligence Review, Jg. 6/ Nr. 1, 35–66.

DSGVO – Datenschutz-Grundverordnung (2003): Erwägungsgrund 13 DSGVO – Berücksichtigung von Kleinstunternehmen sowie kleinen und mittleren Unternehmen, in: DSGVO-

Gesetz (Blog), URL: https://dsgvo-gesetz.de/erwaegungsgruende/nr-13/ (aufgerufen am: 30/03/2025).

Dukino, C. / Ganz, W. / Hämmerle, M. / Renner, T. / Friedrich, M. / Kötter, F. / Meiren, T. / Neuhüttler, J. / Schuler, S. / Zaiser, H. (2019): Künstliche Intelligenz in der Unternehmenspraxis: Studie zu Auswirkungen auf Dienstleistung und Produktion, in: Bauer, W. (Hrsg.): Fraunhofer-Institut für Arbeitswirtschaft und Organisation, Stuttgart: Fraunhofer Verlag.

Europäische Kommission (2003): Empfehlung der Kommission vom 06. Mai 2003 betreffend die Definition der Kleinstunternehmen sowie der kleinen und mittleren Unternehmen, URL: https://eur-lex.europa.eu/legal-content/DE/TXT/PDF/?uri=CELEX:32003H0361 (aufgerufen am: 29/03/2025).

Föcking, J. (2024): Von Schwach bis Super: Die Arten von KI, in: d.velop Blog, URL: https://www.d-velop.de/blog/digitaler-wandel/arten-von-ki/ (aufgerufen am: 29/03/2025).

Franklin, S. / Graesser, A. (1996): Is it an Agent, or Just a Program?: A Taxonomy for Autonomous Agents, in: Müller, J. P. / Wooldridge, M. J. / Jennings, N. R. (Hrsg.): Intelligent Agents III, Agent Theories, Architectures, and Languages, Berlin: Springer, 21–35.

Future Commerce (2023): The State of GenAI x Commerce, URL: https://pages.futurecommerce.fm/hubfs/Reports/Future%20Commerce%20GenAI%20x%20Commerce%20Report.pdf (aufgerufen am: 15/04/2025).

Goertzel, B. / Pennachin, C. (2007): Artificial General Intelligence, New York: Springer.

Heiser, A. (2024): Texten mit ChatGPT. Einfach schnell und kreativ: Ideenmaschine für Kommunikation, Marketing, Werbung und PR, Wiesbaden: Springer.

Körber, A. (2025): Persönliches Interview über die Implementierung des internen KI-Assistenten der Otto Group, geführt am 24.02.2025 [unveröffentliches Interview].

Lederer, M. / Daus, L. (2021): KI in E-Commerce-Prozessen der Modebranche, in: Terstiege, M. (Hrsg.): KI in Marketing & Sales – Erfolgsmodelle aus Forschung und Praxis: Konzepte und Instrumente zum erfolgreichen Einsatz künstlicher Intelligenz, Wiesbaden: Springer, 293–313.

Manhart, K. (2017): Eine kleine Geschichte der künstlichen Intelligenz, in: Computerwoche, URL: https://www.computerwoche.de/article/2756938/eine-kleine-geschichte-der-kuenstlichen-intelligenz.html (aufgerufen am: 29/03/2025).

mrbullwinkle (2025): What is Azure OpenAI Service? – Azure AI Services, URL: https://learn.microsoft.com/en-us/azure/ai-services/openai/overview (aufgerufen am: 29/03/2025).

Otto Group Geschäftsbericht (2023/24): Kennzahlen Otto Group, URL: https://www.ottogroup.com/de/ueber-uns/kennzahlen.php (aufgerufen am: 27/03/2025).

Rancic Moogk, D. (2012): Minimum Viable Product and the Importance of Experimentation in Technology Startups, in: Technology Innovation Management Review, Jg. 2 / Nr. 2, 23–26.

Ribeiro, M. T. / Singh, S. / Guestrin, C. (2016): Model-Agnostic Interpretability of Machine Learning, in: arXiv preprint arXiv:1606.05386, 91–95.

Schurr, P. (2025): Künstliche Intelligenz, in: mindsquare (Blog), URL: https://mindsquare.de/knowhow/kuenstliche-intelligenz/ (aufgerufen am: 30/03/2025).

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

Shortliffe, E. H. / Davis, R. / Axline, S. G. / Buchanan, B. G. / Green, C. C. / Cohen, S. N. (1975): Computer-based Consultations in Clinical Therapeutics: Explanation and Rule Acquisition Capabilities of the MYCIN System, in: Computers and Biomedical Research, Jg. 8 / Nr. 4, 303–320.

Stanford (2012): CS221, URL: https://stanford.edu/~cpiech/cs221/apps/deepBlue.html (aufgerufen am: 29/03/2025).

Statistisches Bundesamt (2025): 55% in kleinen und mittleren Unternehmen tätig, URL: https://www.destatis.de/DE/Themen/Branchen-Unternehmen/Unternehmen/Kleine-Unternehmen-Mittlere-Unternehmen/aktuell-beschaeftigte.html (aufgerufen am: 27/03/2025).

Stowasser, S. / Suchy, O. / Huchler, N. / Müller, N. (2020): Einführung von KI-Systemen in Unternehmen. Gestaltungsansätze für das Change-Management, URL: https://www.isf-muenchen.de/wp-content/uploads/2020/11/AG2_Whitepaper_Change_Management.pdf (aufgerufen am: 29/03/2025).

Teubner, T. / Flath, C. M. / Weinhardt, C. / Aalst, W. van der / Hinz, O. (2023): Welcome to the Era of ChatGPT et al. The Prospects of Large Language Models, in: Business & Information Systems Engineering, Jg. 65 / Nr. 2, 96–101.

Wellner, H. (2025): Persönliches Interview über KI im Einsatz im Modehaus Wellner, geführt am 06.03.2025 [unveröffentliches Interview].

Verantwortung im digitalen Diskurs[1]

Jonah Dähne

1. Einleitung

US-Präsidentschaftswahlkampf 2024: Donald Trump behauptet in einem TV-Duell gegen Kamala Harris, Migranten aus Haiti würden in Springfield, Ohio Haustiere essen. Diese Aussage war jedoch – was der TV-Moderator noch während der Sendung erklärte – komplett falsch, Teil einer Desinformationskampagne, die vor allem über die Social Media Plattform X verbreitet wurde. Dennoch entschied sich Trump im September 2024 vor einem Live-TV-Publikum von über 67 Millionen Menschen, diese bizarre Aussage zu wiederholen. Die Folgen waren fatal: Schulen und Krankenhäuser erhielten in den nächsten Wochen zahlreiche Bombendrohungen, der Rassismus gegen Haitianer nahm stark zu und Neo-Nazi Gruppen wie die Proud Boys organisierten Märsche durch Springfield (vgl. Yousef 2024).

Dieses Phänomen gezielter Verbreitung von Desinformation und Hate Speech im Internet ist nichts Neues, allerdings kommt es durch die zunehmende Deregulierung von Social Media-Konzernen und die Incentivierung von Clickbait Headlines zu immer mehr Vorfällen. Nicht nur in den USA, sondern auch in Deutschland und Europa nehmen Desinformationskampagnen und Hassrede-Vorfälle gegen Einzelpersonen, Bevölkerungsgruppen und Unternehmen zu. Laut einer Studie von Daniela Stockmann und Sophia Schlosser im Auftrag der Amadeu-Antonio Stiftung und des Bundesverbands der Kommunikatoren (BdKom), „nehmen Kommunikatorinnen und Kommunikatoren in Deutschland insbesondere seit dem Jahr 2015 eine steigende Zahl von Hassrede-Fällen

[1] Workshop auf der #CDRK24 Konferenz unter dem Titel „Desinformation und Hate Speech: Welche Rolle haben Unternehmen bei der Gestaltung eines fairen gesellschaftlichen Diskurses?“ – geleitet von Maria Hinz (BARMER), Deniz Taskiran (Telefónica) und Dr. Frank Esselmann (CDR-Initiative), betreut von Jonas Dähne (Universität Bayreuth) und Julius Meier (Universität Bayreuth). Der nachfolgende Beitrag ist im Rahmen eines Blockseminars „Ethik der Digitalisierung“ an der Universität Bayreuth entstanden und wurde durch den Workshop auf der #CDRK24 Konferenz inspiriert.

wahr [...] Im Rahmen ihrer beruflichen Tätigkeit konkrete Erfahrungen mit Hassrede gesammelt haben im privaten Sektor knapp 33 Prozent, im öffentlichen Sektor 59 Prozent und bei NGOs 55 Prozent“ (Stockmann et al. 2019). Hate Speech und Desinformation im digitalen Diskurs gehören also offenbar längst zum Alltag. Für Behörden, Institutionen und Unternehmen stellt sich daher die Frage: Wie sollen sie mit Desinformation umgehen und auf Hassreden reagieren? Können sie es sich leisten, untätig zu bleiben, oder haben sie eine Verantwortung, sich aktiv an der Gestaltung eines besseren öffentlichen Diskurses zu beteiligen? Das Internet ist für viele Menschen längst zu einer der wichtigsten Informations- und Nachrichtenquellen geworden. Was Information und was Desinformation ist, lässt sich für Laien auf den ersten Blick oft nicht mehr erkennen.

Mit diesem schnell wachsenden Problem mussten sich auch viele Telekommunikationsunternehmen Anfang 2020 auseinandersetzen, als zu Beginn der Corona Pandemie mehrfach Desinformation im Internet verbreitet wurde. Es wurde behauptet 5G-Masten könnten möglicherweise Covid19 auslösen oder zumindest würden ihre Frequenzen dazu beitragen das Immunsystem zu schwächen. Daraufhin kam es besonders in Europa und Kanada zu einer Serie an Brandanschlägen auf Sendemasten, Mitarbeiter der Unternehmen wurden vermehrt körperlich und verbal angegriffen. Es ging sogar so weit, dass Unternehmen wie O_2, Vodafone und die kanadische und englische Regierung in öffentlichen Statements erklären mussten, dass 5G nicht Corona auslöse und es sich hierbei um eine Desinformations-Kampagne handelte. Fälle wie diese häufen sich und zeigen, dass Hassreden und Desinformationen „eine ernste Bedrohung für die Qualität des öffentlichen Diskurses in der demokratischen Landschaft Deutschlands darstellt“ (s.o. Stockmann). Um gegen dieses Problem vorzugehen, sollten auch Unternehmen in der Verantwortung stehen, bei der Gestaltung eines transparenten, faktenbasierten öffentlichen Diskurses mitzuwirken.

Im folgenden Text werde ich diese These genauer erklären und untersuchen, wie Unternehmen konkret im Kampf gegen Desinformation und Hate Speech vorgehen können. Zunächst werden im ersten Kapitel die Begriffe Desinformation und Hate Speech definiert und erklärt, wie diese sich von anderen Begriffen abgrenzen lassen. Außerdem werden die negativen Effekte eines digitalen Diskurses, der durch Desinformation und Hate Speech geprägt ist, analysiert und anhand dieser die Relevanz für Unternehmen unterstrichen. Im Anschluss folgt das zweite Kapitel mit dem ersten thematischen Schwerpunkt, Jürgen Habermas’ Diskursethik und wie diese als ethischer Leitfaden für Unternehmen bei der Gestaltung eines besseren digitalen

Diskurses verwendet werden kann. Danach erfolgt im dritten Kapitel unter Einbezug von aktuellen Corporate Digital Responsibility-Praxisbeispielen, eine Analyse interner und externer Handlungsmaßnahmen für Unternehmen; abschließend folgt eine kritische Reflexion und ein Ausblick in die Zukunft.

2. Desinformation und Hate Speech: Gefahr für den digitalen Diskurs

Desinformation und Hate Speech sind häufig verwendete Begriffe, deren Nutzung allerdings oft zu Missverständnissen führt, da viele sie gerne als Sammelbegriff für Fehlinformation und diskriminierende Inhalte verwenden. Unter Desinformation versteht man allerdings laut der Stiftung Wissenschaft und Politik konkret „die gezielte Verbreitung von falschen oder irreführenden Informationen mit der Absicht, einzelne Menschen oder Gruppen zu manipulieren" (Kleemann 2024: 1). Das bedeutet: Versehentliche Falschmeldungen oder Satire mit zugespitzter humoristischer Darstellung fallen nicht unter den Begriff Desinformation. Die Bundesregierung erklärt, der wesentliche Unterschied zwischen einer Fehlinformation und Desinformation sei die gezielte Täuschungsabsicht, die hinter Desinformation steht (vgl. Bundesregierung 2023). Dieser Aspekt ist auch der Grund, warum Desinformation so gefährlich ist: Es kann ein gezieltes Mittel sein, um Personen oder Unternehmen Schaden zuzufügen und dazu beitragen, die Gesellschaft weiter zu spalten. Der Begriff Hate Speech lässt sich definieren als:

> Gruppenbezogene Menschenfeindlichkeit, die ihren Ausdruck in gewalttätiger Sprache findet. Sie verletzt die Würde und die Rechte von Menschen und kann ganze Gruppen einschüchtern (Amadeu Antonio Stiftung 2022).

Hate Speech wird oft von Rechtsextremen, insbesondere im digitalen Diskurs verwendet, um abwertende Aussagen und rassistische Rhetorik zu normalisieren und weiter in der Gesellschaft zu verbreiten. Teil des Problems ist das Phänomen, dass bei Hate Speech auf Social Media Plattformen eine verstärkende Wirkung durch Algorithmen besteht. Inhalte mit Hate Speech führen oft zu starken Reaktionen, entweder Empörung und Wut oder Zustimmung. Das registriert der Algorithmus als hohe Interaktion und „entscheidet" dann, dass der Beitrag relevant ist. Dies führt dazu, dass mehr Leute diese Inhalte sehen und dadurch die Sichtbarkeit und Reichweite der Hate Speech verstärkt wird. Dieses Phänomen konnte man besonders gut am Beispiel der Social Media

Plattform Twitter, heute X, beobachten. Kurz nach der Übernahme von Twitter durch den selbsternannten „Free Speech Absolutist" Elon Musk kam es in den folgenden Monaten zu einem starken Zuwachs (50 Prozent wöchentlich) von Hate Speech auf der Plattform (vgl. Manke 2025).

Der kontinuierlich wachsende Trend von Desinformation und Hate Speech auf Social Media Plattformen ist allerdings nicht nur auf politische und gesellschaftliche Diskurse beschränkt. Jedes Unternehmen, das digitale Kommunikationskanäle nutzt, sei es direkt durch Markenkommunikation und Kundenservice oder indirekt in Form von Werbung, sind Teil des digitalen Diskursraums. Desinformation und Hate Speech vergiften diesen Diskurs, indem beispielsweise gezielt falsche Inhalte geteilt werden. Das kann für Unternehmen negative Auswirkungen haben, da es zu einem Vertrauensverlust der Kunden kommen kann. Beispiele dafür sind die bereits genannte 5G Desinformationskampagne oder gezielte Kampagnen gegen die Corona Impfung. Durch die gezielte Verbreitung von Desinformationen entsteht eine Misstrauenskultur, in der sogar sachlich korrekte Aussagen von Unternehmen in Frage gestellt werden. Aufgrund der zunehmenden Polarisierung im digitalen Diskurs kommt es außerdem häufiger zu gezielten Angriffen auf Mitarbeiter, die selbst Opfer von Hetzkampagnen und Hate Speech werden können.

Auch anhand ökonomischer Sichtweisen lässt sich argumentieren, dass Unternehmen bei der Gestaltung eines ‚gesünderen' Diskurses Verantwortung tragen sollten. Das Konzept der Kooperationsökonomie besagt, dass es in der Wirtschaft nicht nur um Profitmaximierung geht, sondern, dass Zusammenarbeit, Vertrauen und gemeinsame Verantwortung wichtige Bausteine einer nachhaltigen Wirtschaft sind. Eine durch Hate Speech und Desinformation gespaltene Gesellschaft, die keinen stabilen und inklusiven digitalen Diskurs hat, bietet keine gute Basis, um nachhaltig und erfolgreich zu wirtschaften. Um diese Verantwortung in der Praxis umzusetzen, bietet die Corporate Digital Responsibility Initiative Unternehmen eine Möglichkeit, sich aktiv an der Gestaltung einer besseren digitalen Öffentlichkeit zu beteiligen. Durch das Zusammenschließen von Unternehmen aus unterschiedlichen Branchen und der gemeinsamen Entwicklung ethischer Standards, repräsentiert die CDR-Initiative die zentralen Prinzipien der Kooperationsökonomie. Unternehmen haben die Chance, ihr gesellschaftliches Vertrauen zu stärken und sich positiv am Markt zu differenzieren, indem sie sich klar positionieren für digitale Nachhaltigkeit und einen besseren digitalen Diskurs. Außerdem bietet die CDR-Initiative die Chance zum Austausch mit anderen Unternehmen und die Möglichkeit zur Bündelung von Strategien und Ressourcen im Kampf gegen Desinformation und Hate Speech.

3. Habermas Diskursethik als normativer Rahmen

Die vorangegangenen Überlegungen zeigen, dass Hate Speech und Desinformation nicht nur ein rein gesellschaftliches oder politisches Problem sind, sondern auch wichtige Implikationen für Unternehmen haben. Daher stellt sich die Frage, ob es einen ethischen Rahmen oder eine Grundlage gibt, welche diese Verantwortung der Unternehmen begründen kann oder an denen sich Unternehmen zumindest orientieren könnten. Einen solchen theoretischen Bezugsrahmen liefert möglicherweise die Diskursethik von Jürgen Habermas, die sich mit Bedingungen und Regeln eines fairen rationalen Diskurses beschäftigt.

> Die Diskursethik geht von der Grundannahme aus, dass ein Dialog mit dem Austausch von Argumenten Ausgangspunkt der moralischen Entscheidungsfindung ist. Damit dieser Dialog gelingt, muss er herrschaftsfrei vonstattengehen. Habermas ist sich bewusst, dass ein herrschaftsfreier Diskurs faktisch nicht zu erreichen ist, da in jeder Situation einer der Gesprächspartner in irgendeiner Art und Weise überlegen ist. Diese Überlegenheit kann zum Beispiel rhetorisch sein oder aus einer höheren Bildung resultieren. Es braucht folglich Regeln, damit alle entsprechend am Diskurs partizipieren können (Kretzer 2022).

Die in diesem Zitat angesprochenen Diskursregeln sind in drei Kategorien aufgeteilt:

1. Logisch-semantische Regeln: Widerspruchsfreiheit und klare Begriffsverwendung
2. Dialektische Regeln: Ehrlichkeit und Begründungspflicht
3. Pragmatische Regeln: Chancengleichheit und gewaltfreie Teilnahme

Diese Diskursregeln schaffen Rahmenbedingungen, die laut Habermas einen fairen und guten Diskurs ermöglichen. Habermas formuliert außerdem noch vier sogenannte Geltungsansprüche, die erfüllt werden müssen, damit eine Aussage im Diskurs als legitim angesehen wird:

1. Verständlichkeit: Die Aussage muss klar nachvollziehbar sein.
2. Wahrheit: Die Aussage muss mit der objektiven Realität übereinstimmen.
3. Normative Richtigkeit: Die Behauptung muss mit den geltenden sozialen Normen und Werten übereinstimmen.
4. Wahrhaftigkeit: Die Aussagen müssen mit den eigenen Überzeugungen übereinstimmen.

Viele dieser Anforderungen sind im modernen digitalen Diskurs nicht mehr gegeben, was für Habermas ein großes Problem darstellt, da der rationale Diskurs nicht einfach eine idealistische Vorstellung ist, sondern ein wichtiger Teil der Gesellschaft, der zur rationalen Meinungs- und Willensbildung beiträgt. Wer sich an die Diskursregeln hält und die Geltungsansprüche erfüllt, hat die Fähigkeit, Kommunikative Rationalität zu beweisen. Dies passiert, indem man durch verständnisorientierte Kommunikation einen Konsens erreicht. Demgegenüber stellt Habermas die strategische Rationalität als Beispiel dafür, wie ein produktiver Diskurs nicht geführt werden sollte. Habermas beschreibt die strategische Rationalität als Kommunikation, in der man nicht im fairen Austausch zwischen Akteuren einen gemeinsamen Konsens finden will, sondern Kommunikation als Mittel der Durchsetzung eigener Ziele. Strategisches Handeln ist bei Habermas auf Erfolg orientiert; es benutzt die Sprache wie ein Medium, um auf das Verhalten anderer Einfluss zu nehmen (vgl. Habermas 1981: 384 f.).

Umgang mit Hate Speech: Fallbeispiele DAK und Target

Habermas' Vergleich zwischen strategischer und kommunikativer Rationalität lässt sich auch anhand von zwei Beispielen aus dem Unternehmenskontext darstellen. Der Vergleich zeigt die unterschiedlichen Folgen einer Unternehmenskommunikation, die in einem Fall auf kommunikativer und im anderen auf strategischer Rationalität beruht. Anfang 2018 startet die Krankenkasse DAK eine Werbekampagne zum Thema Schwangerschaft mit Plakaten, die ein junges Paar mit einem Ultraschallbild zeigt. Keiner der Verantwortlichen bei der DAK erwartet die Reaktion, die in den Wochen nach der Veröffentlichung folgt. Das Werbeplakat löst im Internet einen rechten Shitstorm aus, weil der werdende Vater auf dem Plakat schwarz ist. Die DAK wird Ziel tausender hasserfüllter Nachrichten, die Werbeplakate werden mit rassistischen Slogans beschmiert und die AFD Nordwestmecklenburg kommentiert das Plakat auf Facebook mit den Worten, es zeige die „Flutung unseres Landes mit Migranten".

Die Kommunikationsabteilung der DAK hat hiermit überhaupt nicht gerechnet und muss schnell handeln. Die Verantwortlichen entscheiden sich für eine einfache, klare, authentische Positionierung gegen die Reaktion. Es stellen sich acht Mitarbeiter auf einem Foto vor das Plakat mit der Nachricht: „Ein junges Paar freut sich auf sein Baby – und wir werden für dieses Motiv kritisiert […] Wir sind eine Krankenkasse für alle Menschen […] Zeigt auch ihr #haltung!" (Kornmann 2020). Dieses Statement erfüllt alle Geltungsansprüche, indem die DAK klar

und transparent ihre Werte verteidigt und sich nicht defensiv und eingeschüchtert vor der potenziellen Reaktion auf dieses Statement verhält. Außerdem fördert die DAK den herrschaftsfreien Diskurs über Toleranz und gegen Rassismus dadurch, dass sie andere Menschen dazu auffordert, ebenfalls Haltung zu zeigen. Der Vorstandschef Andreas Storm kommuniziert darüber hinaus die Werte der DAK in einem ARD-Interview, in dem er sagt, dass die Reaktion auf das Plakat weit weg ist von dem, was die Mehrheit der Menschen in Deutschland heute denkt und ruft auch die Gesellschaft weiter auf, Haltung dagegen zu zeigen. Die Kampagne erfährt viel Zuspruch. 90 Prozent aller Reaktionen sind positiv und die DAK erhält für ihre politische Kommunikation den „Politikaward 2018" in der Kategorie Digital Affairs (vgl. Kornmann 2020). Dieses Beispiel zeigt nicht nur einen erfolgreichen Einzelfall im Kampf von Unternehmen gegen Hate Speech, sondern illustriert auch, wie Unternehmen durch kommunikative Rationalität Verantwortung im digitalen Diskurs übernehmen können.

Ein Gegenbeispiel liefert ein Fall aus den USA. Im Mai 2023 bringt der US-amerikanische Discounter „Target", seine jährliche Pride-Kollektion, eine Reihe von LGBTQ+ freundlichen Produkten, heraus. Diese sollten, so „Target", Vielfalt und Inklusion fördern. Kurz darauf kommt es in verschiedenen Social-Media-Kanälen zu einer Desinformations- und Hass-Kampagne, die behauptet, die Kollektion würde Kinder sexualisieren und eine „radikale LGBTQ+-Agenda" unterstützen. Dies führt dazu, dass Produkte teilweise beschädigt werden. Mitarbeiter in den Filialen werden von Kunden verbal attackiert. Target reagiert, indem einige der umstrittenen Produkte aus ihrem Sortiment entfernt werden. Die Konzern-Verantwortlichen begründen dies mit Sicherheitsbedenken. Zusätzlich veröffentlichen sie ein Statement, dass „Target" weiterhin die LGBTQ+ Gemeinschaft unterstützen würde. Dieses Statement zielt allerdings vor allem auf Schadensbegrenzung. Das Ergebnis ist ein Verlust an Glaubwürdigkeit. Target hatte mit ihrer Pride-Kollektion signalisiert, dass sie eine bestimmte Haltung vertreten, doch kommunizierten dann defensiv und konfliktvermeidend als Reaktion auf externen Druck und die Hass-Kampagne. Anstatt ihre vorgegebenen Werte argumentativ zu verteidigen, entscheidet sich „Target" also, dem Konflikt auszuweichen, was zu dem Vorwurf führte, dass „Targets" Werte offenbar lediglich Image getrieben seien. Diese Gegenüberstellung von zwei unterschiedlichen Kommunikationsstrategien zeigt das mögliche Potenzial eines diskursethischen Rahmens für Unternehmen bei der Reaktion auf Hate Speech und Desinformation.

Habermas Diskursethik im Digitalen Zeitalter

Habermas entwickelte seine Theorien zur Diskursethik in einer Zeit, in der es noch keinen digitalen Diskurs gab. Das bedeutet aber keinesfalls, dass man seine Theorien nicht auf den heutigen Diskurs übertragen kann. In den letzten Jahren hat sich Habermas öfter zum digitalen Strukturwandel geäußert und was dieser für den aktuellen Diskurs und unsere Kommunikation bedeutet. In einem Interview für die Zeitschrift Leviathan anlässlich seines 90. Geburtstags sagte Habermas zur aktuellen Debatte um digitale Diskursräume:

> Die unbestreitbaren Vorteile dieser Technik stellt ja niemand in Frage. Aber im Hinblick auf den Strukturwandel der politischen Öffentlichkeit interessiert mich der eine Aspekt: Sobald die zentrifugalen Kräfte dieser „Blasen" bildenden Kommunikationsstruktur die Sogwirkung der inklusiven Öffentlichkeit aufwiegen, dürften sich konkurrierende öffentliche Meinungen, die für die Bevölkerung im Ganzen repräsentativ sind, nicht mehr ausbilden können. Die digitalen Öffentlichkeiten würden sich dann auf Kosten einer gemeinsamen und diskursiv gefilterten politischen Meinungs- und Willensbildung entwickeln. Soweit ich das heute beurteilen kann, hängt vor allem von der Lösung dieses Problems die Richtung ab, in der sich der Strukturwandel der Öffentlichkeit – und insbesondere der politischen Öffentlichkeit – vollziehen wird (Habermas 2020: 28).

Anhand dieses Zitats kann man erkennen, wie wichtig und zukunftsweisend das Thema rund um die Gestaltung der digitalen Öffentlichkeit und ihrer Diskursräume ist. Habermas betont, dass die Lösung dieses Problems richtungsweisend für den Strukturwandel unserer Öffentlichkeit und der politischen Öffentlichkeit sein wird, somit hängt davon auch die Zukunft unserer Wirtschaft ab. Unternehmen haben in diesem Prozess der Lösungsfindung und Gestaltung im digitalen Diskurs eine große Chance mitzuwirken, um einen nachhaltigen positiven Diskursraum zu schaffen, der die Gesellschaft, Politik und Wirtschaft langfristig stärken wird.

Ein weiterer interessanter Gedanke aus dem Interview ist Habermas' Vergleich von der Erfindung des Buchdrucks und dem Start des Internets. Durch den Buchdruck wurde jeder zum Leser, das Internet hat dagegen jeden zum Autor gemacht. „Der Buchdruck hat alle Nutzer zu „potenziellen Lesern" gemacht, auch wenn es noch drei bis vier Jahrhunderte gedauert hat, bis im

Prinzip alle lesen konnten“ (ebd.: 27). Nun habe das Internet „alle Nutzer zu ‚potenziellen Autoren‘ gemacht – und wie die Nutzer der Presse erst lesen lernen mussten, muss auch die Nutzung des neuen Mediums gelernt werden. Das wird vergleichsweise sehr viel schneller gehen, aber wer weiß, wie lange es dauern wird“ (ebd.). Dieser Vergleich illustriert, dass wir uns noch in der Anfangsphase des digitalen Diskurses befinden und viele Menschen noch lernen müssen, mit diesem neuen Medium umzugehen. Dies kann auch als Anregung gesehen werden für Maßnahmen wie Medienkompetenz-Schulungen, die diesen Lernprozess beschleunigen können.

Abschließend will ich noch ein wichtiges Zitat aus dem Interview hervorheben, in dem Habermas sagt: „Die Diskursethik ist keine Handlungsanweisung, sondern eine Moraltheorie” (ebd.: 14). Dieser Punkt ist wichtig, da Diskursethik und ihre Argumentation sehr theoretisch und idealistisch wirken können und daher, wie Habermas selber sagt, keine genaue Handlungsanweisung zum Beispiel für konkrete Unternehmenskommunikation ist. Dennoch gibt es viele Aspekte der Diskursethik – wie das Konzept der Kommunikativen Rationalität, die Diskursregeln oder die Idee des herrschaftsfreien Diskurses – die einen tragfähigen normativen Rahmen bieten, an dem Unternehmen sich im Kampf gegen Hate Speech und Desinformation orientieren können.

4. Interne und externe Handlungsmöglichkeiten

Das Bewusstsein zu schaffen und Unternehmen zu vermitteln, dass sie als aktive Teilnehmer im öffentlichen Diskurs auch verantwortlich sind für die Gestaltung des digitalen Diskursraums, ist allerdings nur der erste Schritt. Ohne konkrete Maßnahmen und Handlungsmöglichkeiten kann man nicht effektiv gegen Hate Speech und Desinformation vorgehen. Im folgenden Schwerpunktteil werde ich auf konkrete Maßnahmen eingehen, indem ich zuerst interne Handlungsmöglichkeiten beschreibe und dann auf externe Lösungswege eingehe. Eine sehr wichtige und bereits im letzten Absatz erwähnte interne Maßnahme ist die Medienkompetenz-Schulung für Mitarbeiter. Diese kann in Form von Trainings und Workshops stattfinden und Beschäftigten dabei helfen, Desinformation schneller zu erkennen und eine reflektierte Mediennutzung zu entwickeln. Dies baut nicht nur das Fachwissen der Mitarbeiter aus, sondern verbessert auch den digitalen Diskurs, da mehr Leute nun weniger anfällig für Desinformation und andere manipulierende Inhalte sind. Es geht dabei auch um eine Minimierung eines Reputationsrisikos für das Unternehmen: Wenn Mitarbeiter besser informiert sind, ist es weniger wahrscheinlich, dass aus dem

Unternehmen heraus unwissend Desinformation verbreitet wird, was der Firma potenziell schaden könnte. Auch aus der Perspektive der Diskursethik ist dies eine sehr sinnvolle Maßnahme, da durch Medienkompetenz- Training die kommunikative Rationalität der Mitarbeiter gefördert wird. Außerdem gleicht Medienkompetenz-Training die Informationsasymmetrie aus, die zwischen dem Akteur besteht, der bewusst Desinformation verbreitet und dem Rezipienten, der diese unwissend weiterverbreitet. Gleichzeitig wird dadurch auch Habermas Diskursregel der Chancengleichheit befolgt, indem mehr Menschen sich frei von gezielter Manipulation am digitalen Diskurs beteiligen können.

Eine weitere Handlungsmaßnahme ist der Aufbau von internen Krisenteams und Strukturen, die im Fall einer Desinformationskampagne oder wenn Beschäftigte Ziel von Hate Speech geworden sind, schnell reagieren können. Diese Teams können präventiv genaue Abläufe und Strategien entwerfen, die sofort in Kraft treten, sobald sich ein Vorfall ereignet. Dadurch können sich Unternehmen besser auf potenzielle Kampagnen vorbereiten, anstatt erst im akuten Krisenfall zu reagieren. Außerdem sind diese Krisenteams in der Lage, aktiv zur Verbesserung des Diskursraums beizutragen, indem sie konsequent gegen Hate Speech und Desinformation vorgehen. Ein solches internes Krisenteam gibt es bereits bei der Deutschen Telekom: „innerhalb des Unternehmens engagiert sich die Deutsche Telekom, wenn es zu Fällen von Hass und Ausgrenzung kommt. Deshalb hat sie einen internen Hub gegründet, in dem alle Unternehmensbereiche vertreten sind, die bei Vorfällen von Ausgrenzung und Diskriminierung innerhalb der Magenta Welt involviert werden" (Brink et al. 2023: 44). Dieses Beispiel unterstreicht die interdisziplinäre Zusammenarbeit in den Krisenteams, die zu einem effektiven Austausch an Wissen führt und eine schnelle einheitliche Reaktion ermöglicht.

Eine dritte interne Maßnahme im Kampf gegen Desinformation und Hate Speech sind klare Kommunikationsrichtlinien und Community Management. Dazu gehört die Aufstellung von Moderationsteams, die Unternehmenskanäle auf Kommentare, die Hate Speech und Desinformation enthalten prüft und diese dann löscht. Das führt dazu, dass ein fairer Diskurs durch Moderationsteams und Kommunikationsrichtlinien gesichert werden kann. Hier gibt es einen direkten Bezug zu Habermas Geltungsansprüchen. Unternehmen können Aussagen anhand von diskursethischen Geltungsansprüchen wie Wahrheit und normativer Richtigkeit prüfen und diese entfernen, wenn sie den Geltungsansprüchen nicht entsprechen. Darüber hinaus fördern Moderationsteams und Kommunikationsrichtlinien die Durchsetzung der Diskursregel der gewaltfreien Teilnahme, indem sie Hate Speech im digitalen Diskurs entfernen, die es darauf abzielt durch Einschüchterung

und Abgrenzung Personen oder Gruppen vom Diskurs auszuschließen. Gleichzeitig wird durch Entfernen von Desinformation und Hate Speech Habermas Konzept der kommunikativen Rationalität gefördert und die strategische Rationalität, die hinter Hate Speech und Desinformation steckt, bekämpft. Ein weiterer Aspekt dieser Maßnahmen sind klare Leitlinien für die öffentliche Kommunikation. Ein Beispiel für die erfolgreiche Umsetzung solcher Maßnahmen ist O_2 Telefonica, die Mitarbeiter im Kundenservice zu speziellen Schulungen schickt, damit sie lernen, wie sie professionell mit Nutzern kommunizieren und in kritischen Situationen richtig reagieren können.

Externe Maßnahmen

Eine der wichtigsten externen Maßnahmen zur erfolgreichen Gestaltung eines besseren digitalen Diskurses ist die Förderung von Bildung und digitaler Aufklärung. Ähnlich wie bei einer internen Medien-Kompetenz Schulung, aber im deutlich größeren Maßstab können Unternehmen die Gesellschaft gegen Desinformation und Hate Speech stärken, wenn sie in die Förderung von Bildung und digitaler Aufklärung investieren. Die Deutsche Telekom hat bereits Konzepte speziell für Kinder, Eltern und Lehrer entwickelt. Ein Beispiel ist die Initiative „teachtoday.de", die die Medienkompetenz von Kindern und Jugendlichen fördern soll. „Teachtoday" bietet Materialien und konkrete Handlungsempfehlungen für Lehrer und Eltern (vgl. Telekom 2024).

> Hier wird deutlich gemacht, wie sich Online- und Offline-Welt unterscheiden, welche Auswirkungen Hassrede und Meinungsmache haben und was jede*r konkret dagegen tun kann – von der direkten Gegenrede bis zur Anzeige beim Plattformbetreiber oder bei der Polizei. Über die Plattformen, aber auch in Workshops, über Web- und Podcasts wurden allein 2022 vier Millionen Menschen erreicht" (Brink et al. 2023: 44).

Solche Maßnahmen bieten wichtige zukunftsorientierte Handlungsmöglichkeiten, um jüngere Generationen auf den Umgang mit digitalen Diskurs Räumen vorzubereiten. Außerdem fördern diese Bildungsinitiativen ähnlich wie internes Medienkompetenz-Training die Diskursfähigkeit der Gesellschaft. Teilnehmer lernen, wie sie strategisch verzerrte Kommunikation im digitalen Diskurs erkennen können und wie ein verbesserter digitaler Diskursraum gestaltet werden kann. Maßnahmen wie diese unterstützen zusätzlich die Umsetzung der Diskursregel der Chancen-

gleichheit, da durch Bildung mehr Menschen sich frei von Täuschung am digitalen Diskurs beteiligen können. Bildungsinitiativen fördern außerdem die Sicherung des Geltungsanspruches der Wahrheit, weil durch Aufklärung das Erkennen von Desinformation erleichtert und unbewusste Verbreitung von Falschinformation reduziert wird.

Die Förderung von Bildung und digitaler Aufklärung kann Krisen in Bereichen verhindern, die besonders anfällig für Desinformation sind. Die Barmer Krankenkasse setzte während der Corona Pandemie beispielsweise auf frühzeitige Aufklärung und positionierte sich als vertrauenswürdige Informationsquelle.

> Begleitend dazu klärte die BARMER in aktuellen qualitativ gesicherten Beiträgen und Faktenchecks crossmedial rund um die Pandemie auf. Studien und Umfragen unterstützten dabei, ein klareres Bild über die Situation zu gewinnen und diese im Sinne der Menschen durch gesicherte Informationen positiv zu beeinflussen. In der aufgeheizten Debatte sollten qualitätsgesicherte Informationen und konkrete Handlungsempfehlungen den Menschen Sicherheit und Orientierung geben (ebd.: 40).

Dieses Beispiel zeigt, wie ein Unternehmen während einer Krise, die stark von Desinformation geprägt war, gesellschaftliche Verantwortung übernimmt und sich für faktenbasierte Argumentation einsetzt und somit Haltung zu seinen Werten beweist.

Die verantwortungsbewusste Steuerung von Werbebudgets ist eine zusätzliche externe Maßnahme zur Eindämmung von Desinformation. Wichtig ist dabei, einen kritischen Blick auf die Lieferkette zu werfen, die hinter Desinformation und Hate Speech steht. Das Ziel einer solchen Analyse ist es, Werbeplatzierung auf Plattformen oder Seiten mit viel Desinformation und Hate Speech zu vermeiden. Durch diese gezielte Werbestrategie vermeiden Unternehmen, mit diesen Inhalten in Verbindung gebracht zu werden und stärken diskursethische Werte im Markt. Außerdem entziehen sie so Seiten oder Plattformen, die gefährliche Inhalte verbreiten, ihre finanzielle Unterstützung. Zur Umsetzung einer verantwortungsbewussten Werbeplatzierung können Unternehmen Brand Safety Tools verwenden, die verhindern, dass ihre Werbung auf Seiten oder Plattformen mit Hate Speech oder Desinformation geschaltet wird. Dies ist ein besonders wichtiger Schritt, da ein Wegfall von Werbegeldern möglicherweise Seiten oder Plattformen dazu bewegt mehr Moderation und Content Checks

durchzuführen, um problematische Inhalte zu beseitigen. Letztlich ist eine der effektivsten Maßnahmen, die Werte des Unternehmens klar und transparent zu kommunizieren und Haltung zu beweisen, wenn diese durch Desinformation und Hate Speech bedroht werden. Diese Maßnahme repräsentiert außerdem direkt Habermas diskursethischen Geltungsanspruch der Wahrhaftigkeit als Kriterium für einen fairen Diskurs, da die Aussagen eines Unternehmens mit dessen Überzeugungen übereinstimmen müssen. Unternehmen haben hier die Möglichkeit, in einem zunehmend toxischen digitalen Diskurs durch Kampagnen wie ‚Gegen Hass im Netz' oder ‚Haltung' eine Vorbildfunktion einzunehmen und Diskurse zu verschieben, weg vom Hass, hin zu mehr Toleranz und Inklusion.

5. Reflexion und Ausblick

Dies führt zurück zur Leitfrage und Kernaussage dieses Textes. Ein Unternehmen hat als Teilnehmer am digitalen Diskurs die Chance und die Verantwortung, aktiv an der Gestaltung einer besseren öffentlichen Kommunikation mitzuwirken und sollte dies auch tun. Denn ein fairer, transparenter digitaler Diskurs ist eine gute Basis für eine funktionierende demokratische Gesellschaft und Wirtschaft. Dabei ist ein verantwortungsvoller Umgang mit digitaler Kommunikation und ein Engagement im digitalen Raum eine Win-Win-Situation. Denn aufgeklärte Verbraucher und Kunden, die von Unternehmen auf Augenhöhe angesprochen werden, tragen langfristig zum Unternehmenserfolg bei. Allerdings bleibt die Bekämpfung von Desinformation auch in Zukunft eine große Herausforderung. Denn aufgrund der sich konstant weiterentwickelnden technischen Mittel wie KI oder Deep Fakes, wird die Verbreitung von Desinformation immer effektiver, während gleichzeitig die Detektion und Bekämpfung immer schwerer und aufwendiger werden. Ich kann mich daher nur Jürgen Habermas' Worten anschließen, dass die Lösung dieses Problems richtungsweisend für den Strukturwandel unserer und der politischen Öffentlichkeit sein wird und somit auch immensen Einfluss auf die Zukunft unserer Wirtschaft haben wird. In turbulenten Zeiten, in denen die Gesellschaft immer mehr durch Polarisierung in digitalen Räumen gespalten wird, ist es wichtig, dass Unternehmen geschlossen und konsequent gegen die Vergiftung des digitalen Diskurses vorgehen. Hier bieten Initiativen wie Corporate Digital Responsibility einen guten Rahmen für die gemeinsame Bekämpfung von Hate Speech und Desinformation. Abschließend lässt sich sagen, dass konkrete Handlungsmöglichkeiten in Verbindung mit einem normativen Rahmen wie der Diskursethik von Jürgen Habermas Unternehmen eine effektive Basis bieten können, um Desinformation und Hate Speech wirksam zu bekämpfen.

Amadeu Antonio Stiftung (2025): Was ist Hate Speech?, URL: https://www.amadeu-antonio-stiftung.de/digitale-zivilgesellschaft/was-ist-hate-speech/ (aufgerufen am: 05/04/2025).

Berner, N. (2024): Wer hat, dem wird gegeben – Über die Logik von gesellschaftspolitischen Diskursen auf sozialen Medien, Bundeszentrale für Politische Bildung, URL: https://www.bpb.de/themen/medien-journalismus/soziale-medien/545478/wer-hat-dem-wird-gegeben-ueber-die-logik-von-gesellschaftspolitischen-diskursen-auf-sozialen-medien/ (aufgerufen am: 07/04/2025).

Brink, A. / Esselmann, F. / Hinz, M, Mehlmann-Tripp, M. (2023): Klare Kante gegen Desinformation und Hate Speech! Wie Unternehmen Verantwortung übernehmen und dabei auch ihr Geschäft schützen, herausgegeben von der Geschäftsstelle CDR-Initiative, URL: https://cdr-initiative.de/uploads/files/CDR_Initiative_WP_Desinformation_Hate_Speech.pdf (aufgerufen am: 01/04/2025).

Bundesregierung (2023): Begriffsklärung: Was ist Desinformation?, URL: https://www.bundesregierung.de/breg-de/aktuelles/was-ist-desinformation-1875148 (aufgerufen am: 05/04/2025).

CDR-Initiative (2024): Corporate Digital Resonsibility-Kodex, Freiwillige Selbstverpflichtung mit Bericht, Stand: Juni 2021, URL: https://cdr-initiative.de/uploads/files/2024-01_Kodex_CDR-Initiative.pdf (aufgerufen am: 01/04/2025).

El Ouassil, S. (2021): Habermas und die Demokratie 2.0, URL: https://www.spiegel.de/kultur/juergen-habermas-strukturwandel-der-oeffentlichkeit-in-der-2-0-version-a-2e683f52-3ccd-4985-a750-5e1a1823ad08 (aufgerufen am: 28/03/2025).

Habermas, J. (2020): Moralischer Universalismus in Zeiten politischer Regression, Jürgen Habermas im Gespräch über die Gegenwart und sein Lebenswerk, in: Leviathan, Jg. 48 / Nr. 1, 7–28, URL: https://www.nomos-elibrary.de/de/10.5771/0340-0425-2020-1-7.pdf?download_full_pdf=1&page=1 (aufgerufen am: 03/04/2025).

Habermas, J. (1991): Erläuterungen zur Diskursethik, Frankfurt am Main: Suhrkamp.

Habermas, J. (1981): Theorie des kommunikativen Handelns, Band 1: Handlungsrationalität und gesellschaftliche Rationalisierung, Band 2: Zur Kritik der funktionalistischen Vernunft, Frankfurt am Main: Suhrkamp.

Kleemann, A. (2024): Wie man erfolgreich Desinformation bekämpft, Stiftung Wissenschaft und Politik, URL: https://www.swp-berlin.org/publications/products/aktuell/2024A69_Desinformation_Bekaempfung.pdf (aufgerufen am: 04/04/2025).

Kornmann, J. (2020): #Haltung: Wie Die Dak-Gesundheit Einen Rechten Shitstorm Pariert Hat: Hate Speech. Hate Speech | Ein Ratgeber für Kommunikationsverantwortliche, URL: https://www.hatespeechentgegentreten.de/haltung-gegen-hass-und-hetze-krankenkassenwerbung-ein-rechter-shitstorm-und-die-reaktion-der-dak-gesundheit/ (aufgerufen am: 06/04/2025).

Kretzer, C. (2022): Philosoph des Monats: Jürgen Habermas, Reli-Ethik-Blog, Religion, Ethik und Philosophie bei Klett, URL: https://reli-ethik-blog.de/philosoph-des-monats-juergen-habermas/ (aufgerufen am: 07/04/2025).

Manke, K. (2025): Study Finds Persistent Spike in Hate Speech on X, Berkeley News vom 13.02.2025, URL: https://news.berkeley.edu/2025/02/13/study-finds-persistent-spike-in-hate-speech-on-x/ (aufgerufen am: 05/04/2025).

Rat für Nachhaltige Entwicklung (2024): eXit Hate Speech: 47 Organisationen stellen Aktivitäten auf X ein, URL: https://www.nachhaltigkeitsrat.de/aktuelles/exit-hate-speech-47-organisationen-stellen-aktivitaeten-auf-x-ein/ (aufgerufen am: 08/04/2025).

Schloemann, J. (2020): Habermas und das Internet, Lernen, lernen, lernen, in: Süddeutsche Zeitung vom 12.03.2020, URL: https://www.sueddeutsche.de/kultur/habermas-und-das-internet-lernen-lernen-lernen-1.4841981 (aufgerufen am: 06/04/2025).

Stockmann, D. / Schlosser, S. / School, H. (2019): Wie PR-Experten Hassrede erleben, Studie: Hassrede in der öffentlichen Kommunikation in Deutschland: Wie nehmen PR-Experten sie in ihrer Arbeit wahr?, URL: https://www.hatespeechentgegentreten.de/studie/ (aufgerufen am: 01/04/2025).

Yousef, O. (2024): In Parroting a Lie About Migrants in Springfield, Ohio, Trump Excites Extremists, URL: https://www.npr.org/2024/09/24/nx-s1-5118438/neo-nazi-haitian-springfield-trump-debate (aufgerufen am: 27/04/2025).

Erfolgsfaktoren der AI-Adoption[1]

Praxisnahe Erkenntnisse und organisationale Übertragbarkeit

Leonhard Henke, Alexander Brink und Frank Esselmann

1. Warum AI-Adoption mehr ist als Technologie

Künstliche Intelligenz (KI) transformiert nicht nur einzelne Produkte und Prozesse, sondern wirkt tiefgreifend auf gesamte Organisationen – beispielsweise im Personalwesen durch die automatisierte Sichtung von Bewerbungen und die Vorbereitung von Interviews oder in der Unternehmenssteuerung durch datengetriebene Analysen, die Entscheidungsprozesse fundamental verändern. AI-Adoption bezeichnet nun den konkreten Schritt von der reinen Technologiebeobachtung hin zur praktischen Nutzung von Künstlicher Intelligenz im Unternehmen. Wer darunter jedoch lediglich verkürzt die „technische Implementierung" versteht, verkennt die eigentliche Tragweite. AI-Adoption bedeutet für uns daher mehr, nämlich Künstliche Intelligenz so in Unternehmensprozesse einzuführen, dass diese effizienter, leistungsfähiger und zukunftssicher werden – und zugleich im Einklang mit grundlegenden Prinzipien wie Verantwortung, Transparenz und Fairness stehen. Eine herausfordernde Aufgabe. Denn aufgrund der transformativen und

[1] Die im Beitrag dargestellten Erkenntnisse beruhen auf der Zusammenarbeit im Rahmen der Corporate Digital Responsibility (CDR)-Initiative des BMJV sowie auf Gesprächen und Projekten mit verschiedenen Praxispartnern, darunter: Otto Group, ING-DiBa Deutschland, BARMER, trail GmbH, W. MARKGRAF GmbH & Co KG, Deutsche Kreditbank AG, Telefónica Deutschland, Deutsche Telekom, New Work SE, Atruvia AG, bitkasten GmbH, Zalando SE, AOK PLUS, ALH Gruppe, DigitalSchoolstory gGmbH, Evangelische Landeskirche in Württemberg, Mobil ISC GmbH und HAKRO GmbH. Das generative KI-Tool (OpenAI) wurde in unterstützender Funktion für stilistische Verfeinerungen, sprachliche Optimierungen und als Inspiration für alternative Formulierungen eingesetzt. In wenigen Fällen wurden erste Entwürfe von Passagen mit dem Tool erstellt und anschließend von den Autoren grundlegend überarbeitet und umformuliert. Alle intellektuellen Inhalte, kritischen Überlegungen und endgültigen Entscheidungen liegen in der alleinigen Verantwortung der Autoren.

komplexen Natur von AI verlangt ihre erfolgreiche Adoption nach einem System von Regeln, Prozessen und Beziehungen, durch das Organisationen geführt und kontrolliert werden. Diese lässt sich grundlegend in eine explizit-formal und eine implizit-informell gelebte AI-Governance differenzieren, die jeweils unterschiedliche Reifegrade haben können.

- Explizite Governance (formal): Darunter fassen wir formale Strukturen, Regeln und Verfahren, die – je nach Reifegrad – schriftlich fixiert, transparent und verbindlich sind. Beispiele sind Richtlinien, Kodizes, Compliance, Berichtswege oder festgelegte Verantwortlichkeiten.
- Implizite Governance (informell): Hier geht es um ungeschriebene Regeln, Werte und Praktiken, die im Alltag das Handeln prägen, ohne offiziell dokumentiert zu sein. Dazu gehören geteilte Überzeugungen, kulturelle Normen, Vertrauen zwischen Akteuren oder informelle Netzwerke.

Diese Unterscheidung lässt sich auch aus der Perspektive es US-Organisationspsychologen Edgar H. Schein erklären: Das Modell der Kulturebenen differenziert nämlich in eine explizite Dimension (formal sichtbare Strukturen, Regeln und Verfahren) und eine implizite Dimension (informell gelebte Normen und grundlegende Annahmen). Explizite Governance manifestiert sich – nach unserem Verständnis und an Schein angelegt – in Artefakten und bekundeten Werten (z. B. Richtlinien, Kodizes, Rollen), während die implizite Governance auf den tief verankerten, häufig unbewussten Annahmen beruht, die das tatsächliche Handeln prägen (z. B. Vertrauenslogiken, informelle Netzwerke). Die Wirksamkeit der Steuerung hängt damit wesentlich von ihrer Passung zu den zugrundeliegenden kulturellen Annahmen ab (vgl. Schein/Schein 2017). Während explizite Governance also auf formalisiertem, dokumentiertem Wissen basiert, stützt sich implizite Governance stärker auf erfahrungsgebundenes, schwer kodifizierbares Wissen (vgl. Nonaka/Takeuchi 1995). Unser Reifegradmodell als Wissensspeicher, wie wir es im Fortgang des Beitrags einführen, greift diese Differenzierung von Wissensformen auf, indem es sowohl formalisierbares, dokumentiertes Wissen als auch implizit gelebte, erfahrungsbasierte Praktiken berücksichtigt.

Um diese explizite und implizite Governance nun wirkungsvoll in die Praxis umzusetzen – das verstehen wir unter AI-Adoption –, sind ein starkes Engagement der Führungskräfte sowie eine Kultur der Transparenz und des Lernens notwendig, die Mitarbeitende einbezieht, funktionsübergreifende Zusammenarbeit stärkt und innerorganisationaler Kommunikation hohen Wert beimisst. Ohne in diesen Bereichen Maßnahmen zu treffen, werden Bemühungen zur Einführung von AI auf erheblichen Widerstand innerhalb von Organisationen stoßen. Die Angst vor dem

Verlust des Arbeitsplatzes, mangelndes Vertrauen in AI und Widerstand gegen Veränderungen, z. B. aufgrund von Überforderung sind, dabei häufige Hindernisse.

Erfolgreiche AI-Adoption gelingt nur dann, wenn die Organisation versteht, wo sie steht und was der davon ausgehend nächste sinnvolle Schritt ist. Reifegradmodelle dienen dabei als Strukturrahmen. Sie machen Reife sichtbar, zeigen Entwicklungspfade auf und erlauben passgenaue Entscheidungen statt pauschaler Blaupausen. Das Reifegradmodell lässt sich in diesem Sinne als Wissensspeicher verstehen, der unterschiedliche Wissensformen aufgreift. Geht man mit dieser ganzheitlichen Sicht an die Transformation, kann man das Argument entkräftigen, bei der AI-Adoption ginge es nur um überkomplexe Strukturen und Prozesse auf Kosten von Effizienz. Im Gegenteil: Es geht um die Förderung von Verantwortungsbewusstsein und die Eröffnung von Spielräumen für eine effiziente, zukunftssichere und verantwortungsvolle Nutzung Künstlicher Intelligenz.

2. Das AI-Governance-Projekt der CDR-Initiative

2.1 Das CDR-Reifegradmodell

Die CDR-Initiative, 2018 vom Bundesministerium der Justiz und für Verbraucherschutz ins Leben gerufen, versteht Corporate Digital Responsibility als freiwilliges Engagement von Unternehmen über gesetzliche Vorgaben hinaus, um die digitale Welt im Sinne der Gesellschaft aktiv mitzugestalten (vgl. CDR-Initiative 2025a). Sie fördert die strategische Verankerung digitaler Verantwortung und unterstützt Unternehmen mit Leitfäden, Workshops und Dialogformaten. International gilt sie als Vorreiter (vgl. Merbecks 2023). Ein Meilenstein war der deutsche CDR-Kodex von 2021, der freiwillige Selbstverpflichtungen bündelt (vgl. CDR-Initiative 2025b). In drei klar differenzierten Formen der Zusammenarbeit – Mitgliedschaft, Themenpartnerschaft und Entwicklungspartnerschaft – wird kontinuierlich Wissen erhoben, aufbereitet und im CDR-Reifegradmodell angereichert. Während Mitglieder langfristig und aktiv das Netzwerk mitgestalten, ermöglichen Themenpartnerschaften punktuelle Beiträge zu spezifischen Themen wie Datenethik oder algorithmischer Fairness. Entwicklungspartnerschaften wiederum bieten Organisationen eine enge methodische Begleitung, um CDR systematisch in Strukturen und Prozesse zu integrieren. So schafft die Initiative flexible Zugänge – vom Ideengeber bis zum Umsetzungspartner – und treibt eine werteorientierte digitale Transformation voran (vgl. CDR-Initiative 2025a).

Corporate Digital Responsibility – so wie wir sie rekonstruieren – bietet mit einer klaren Architektur Orientierung für verantwortungsvolle Digitalisierung. Zentrales Element ist das CDR-Reifegradmodell, eingebettet in eine Governance-Struktur, die Steuerung und Weiterentwicklung koordiniert. Aufgebaut ist die CDR-Governance aus sieben ineinandergreifenden Bausteinen (vgl. Brink et al. 2025): den neun CDR-Prinzipien (1) als normatives Fundament, dem CDR-Kodex (2) als freiwilliger Selbstverpflichtung, fünf CDR-Handlungsfeldern (3) mit konkreten CDR-Zielen (4), deren Umsetzung durch CDR-Maßnahmen (5) erfolgt, ergänzt durch den CDR-Quick-Check (6) und schließlich das CDR-Reifegradmodell (7) zur systematischen Bewertung und Weiterentwicklung. In enger Abstimmung zwischen Politik und Wirtschaft ist über mehrere Jahre hinweg ein stimmiges Gesamtkonzept entstanden, in dem die Elemente sinnvoll aufeinander aufbauen.

Im Zentrum des Gestaltungsökosystems für AI-Adoption steht das CDR-Reifegradmodell (RGM) als Wissensspeicher und damit als eine Art Steuerungs- und Lernsystem, das Wissen aus Praxis und Forschung aufnimmt, systematisiert und in Form von Orientierungen und Entwicklungspfaden an Organisationen zurückspielt („Wissensimport" und „Wissensexport"). Es verwendet bewährte und erprobte Formate, Methoden und Zusammenarbeitsformen, um den Status digitaler Verantwortung in Organisationen sichtbar zu machen, Entwicklungspfade aufzuzeigen und CDR mit bestehenden Digital- und Nachhaltigkeitsstrategien zu verknüpfen. Dabei orientiert er sich konsequent an der unternehmerischen Praxis und speist sich aus den Erfahrungen langjähriger Erprobung bei der digitalen Transformation.

Die „Innenhülle" des Reifegradmodells bilden die Gestaltungselemente, die Unternehmen konkrete Instrumente, Impulse und Strukturen für verantwortungsvolle Digitalisierung an die Hand geben. Die „Außenhülle" wird durch Kooperationen gebildet, die den Dialog zwischen Wirtschaft, Politik, Wissenschaft und Zivilgesellschaft ermöglichen und das System nach außen öffnen. So entsteht ein kohärentes Ökosystem, in dem Wissen systematisch gesammelt, strukturiert und über Kooperationen zurück in die Praxis gespielt wird – mit dem Ziel, AI-Adoption verantwortungsvoll und nachhaltig zu gestalten (siehe Abbildung 1).

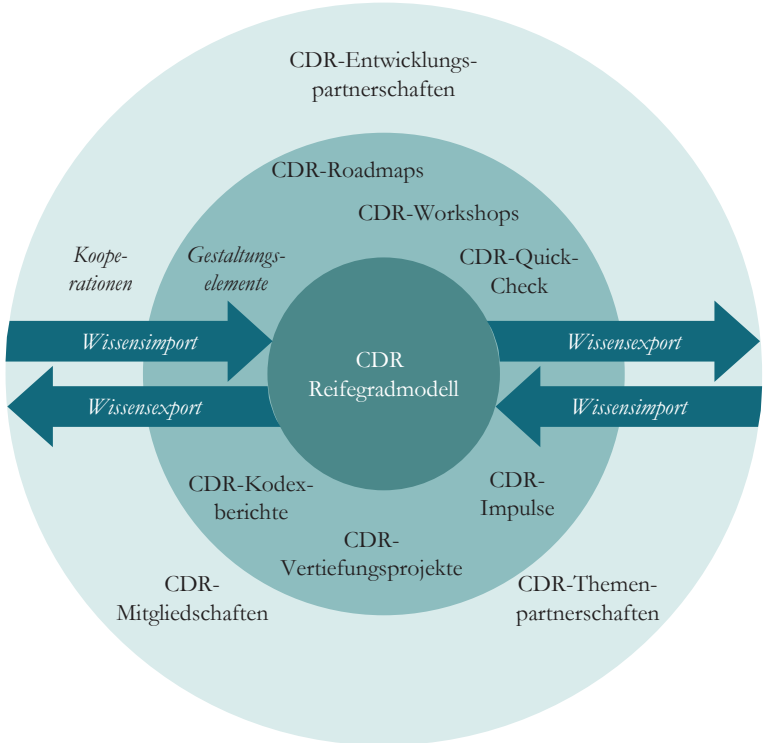


ABBILDUNG 1: DAS CDR-REIFEGRADMODELL ALS WISSENSSPEICHER FÜR DEN STRUKTURIERTEN AUFBAU VON DIGITALVERANTWORTUNG (QUELLE: EIGENE ABBILDUNG)

2.2 Das AI-Governance-Projekt der CDR-Initiative

Im Rahmen eines AI-Governance-Projekts der CDR-Initiative wurden gemeinsam mit den langjährigen Mitgliedern[2] in ihrer Rolle als CDR-Vorreiter zunächst die wichtigsten Fragestellungen bezüglich AI-Governance identifiziert, die sich beim Einsatz von Künstlicher Intelligenz in Organisationen stellen, um so den Fokus auf die wichtigsten Erfolgsfaktoren für AI-Adoption zu

[2] Otto Group, ING-DiBa Deutschland, BARMER, Deutsche Kreditbank AG, Telefónica Deutschland, Deutsche Telekom, W. MARKGRAF GmbH & Co KG.

legen. Um diese Fragen nicht nur auf diese Organisationstypen bezogen beantworten zu können, sondern möglichst praxistaugliche Antworten für einen organisationsübergreifenden Transfer zu entwickeln, werden die gewonnenen Erkenntnisse nunmehr kontinuierlich in Entwicklungspartnerschaften mit unterschiedlichen Organisationen und einhergehenden unterschiedlichen organisationsspezifischen Anforderungen angewendet und weiterentwickelt.

Das CDR-Reifegradmodell schafft dabei eine klare, aber zugleich flexible Struktur, mit der sich digitale Verantwortung – und damit auch AI-Governance – schrittweise aufbauen lässt (vgl. Brink et al. 2025). Wie in Abbildung 2 dargestellt, beschreibt das Modell fünf Entwicklungsstufen organisationaler Strukturen und Prozesse: von ersten, noch unsystematischen Impulsen („Initial“) bis hin zur umfassenden, strategisch verankerten Integration („Optimiert“). Jede Stufe ist durch spezifische Merkmale gekennzeichnet, die den jeweiligen Reifegrad der Organisation sichtbar machen. So werden beispielsweise auf der Stufe „Projektiert“ erste Strukturen erprobt, während auf der Stufe „Definiert“ bereits klare Ziele formuliert und Verantwortlichkeiten festgelegt sind. Mit zunehmendem Reifegrad treten systematische Steuerungselemente hinzu („Gesteuert“) und münden schließlich in einen kontinuierlichen Verbesserungsprozess, bei dem digitale Verantwortung eine zentrale Rolle in der Unternehmensstrategie einnimmt („Optimiert“). Das Modell baut auf dem bekannten Capability Maturity Model Integration (CMMI) auf, wurde jedoch gezielt für die Besonderheiten digitaler Verantwortung im Rahmen der CDR-Initiative weiterentwickelt (vgl. Crosby 2025). Es integriert dabei nicht nur technische und organisatorische, sondern auch kulturelle und ethische Dimensionen.

Reifegrad	Beschreibung
Initial	CDR-Themen werden vereinzelt aufgegriffen, aber noch nicht systematisch bearbeitet.
Projektiert	Erste Strukturen entstehen, die Organisation experimentiert mit CDR-Ansätzen. Temporäre Widersprüche mit anderen Unternehmenszielen sind noch akzeptabel.
Definiert	Die CDR-Ziele sind klar formuliert, Aufgaben sind zugeordnet und unternehmensweit kommuniziert.
Gesteuert	Der Fortschritt wird regelmäßig überprüft, es gibt Kennzahlen zur Erfolgsmessung.
Optimiert	Die CDR-Maßnahmen werden kontinuierlich verbessert und haben eine hohe Priorität in der Unternehmensstrategie.

ABBILDUNG 2: DAS REIFEGRADMODELL
(QUELLE: BRINK ET AL. 2025)

Im Zentrum des CDR-Reifegradmodells stehen die fünf Handlungsfelder des CDR-Kodex – Umgang mit Daten, Bildung, Klima- und Ressourcenschutz, Mitarbeitenden-Einbindung und Inklusion – ergänzt um eine übergreifende Governance-Betrachtung. Nicht jedes Unternehmen soll auf denselben Stand gebracht werden, weil Ausgangslage, Branchenkontext und strategische Prioritäten stark variieren. Ein Start-up mit datengetriebenem Geschäftsmodell etwa steht vor anderen Herausforderungen als ein Industrieunternehmen mit energieintensiver Produktion. Während das eine etwa einen Schwerpunkt auf algorithmische Fairness legen muss, ist für das andere Klima- und Ressourcenschutz besonders relevant.

Das CDR-Reifegradmodell berücksichtigt genau diese Unterschiede: Es zwingt nicht zu einem einheitlichen Idealbild, sondern ermöglicht eine maßgeschneiderte Entwicklung entlang realistischer Ambitionsniveaus. Es erlaubt eine differenzierte Analyse nach sinnvollem Ambitionsniveau: Was ist bereits etabliert? Wo bestehen Lücken, die gefüllt werden sollten? Und was wäre ein sinnvoller nächster realistischer, wirksamer und anschlussfähiger Schritt? Gerade für die Zwillingstransformation ist dieser Abschnitt zentral, weil er zeigt, dass digitale und nachhaltige Entwicklung nur dann wirksam zusammengedacht werden können, wenn Unternehmen differenziert nach ihren Ansprüchen vorgehen und individuelle, aber anschlussfähige Fortschritte machen (vgl. Brink 2022). Entscheidend ist nicht, dass alle denselben Zielzustand erreichen, sondern dass jedes Unternehmen wirksame und anschlussfähige nächste Schritte identifiziert, die sowohl zur eigenen Situation passen als auch zur gemeinsamen Verantwortung für eine nachhaltige digitale Transformation beitragen.

Ein Beispiel ist das Handlungsfeld „Umgang mit Daten“: Ein Unternehmen könnte bereits über eine solide Datenschutzpraxis verfügen, aber noch Lücken bei der algorithmischen Fairness aufweisen. Ein nächster Schritt wäre dann, Transparenz- und Prüfkriterien für KI-Systeme einzuführen. Im Handlungsfeld „Klima- und Ressourcenschutz“ wiederum mag die Einführung energieeffizienter Server-Infrastrukturen bereits umgesetzt sein, während Strategien für eine nachhaltige Software-Entwicklung fehlen. Hier wäre ein pragmatischer Fortschritt, Kriterien für ressourcenschonende IT in Beschaffungs- und Entwicklungsprozessen zu verankern. Diese Logik – Lücken identifizieren, Ambitionsniveau definieren, nächste Schritte gehen – ist im Mitgliedschaftsmodell der Entwicklungspartnerschaften der CDR-Initiative zentral: Unternehmen entwickeln sich nicht nach einem Einheitsmuster, sondern entlang individueller, aber anschlussfähiger Pfade.

3. Mit den wichtigsten Fragen zu den bestmöglichen Antworten

Am Anfang des AI-Governance-Projekts der CDR-Initiative stand ein Katalog von 90 Fragen, aufgebaut auf einschlägigen AI-Governance-Frameworks wie dem NIST AI Risk Management Framework und dem AIGA Framework (vgl. Mäntymäki et al. 2022; National Institute of Standards and Technology 2023). Das Projekt im Rahmen der CDR-Initiative hatte zum Ziel, Unternehmen beim Aufbau einer schlanken, EU-AI-Act-konformen AI-Governance zu unterstützen. In mehreren Runden und mit über 20 Expert*innen der Mitgliedsunternehmen verdichtete die CDR-Initiative diese zu 27 priorisierten Leitfragen. Zwei Themen wurden organisationsunabhängig als prioritär für eine erfolgreiche AI-Adoption im Sinne der Geschäftsentwicklung identifiziert und vertieft. Sie sind hier mit den dazugehörigen priorisierten und diskutierten Fragen dargestellt.

„Kultur & Literacy" bezeichnet die Entwicklung rollenbezogener Kompetenzen, die Etablierung gemeinsamer Narrative und die Förderung von Risiko- sowie Potenzialbewusstsein, um den Umgang mit AI systematisch in die organisatorische Praxis einzubetten. Die erarbeiteten Fragen lauten:

- Inwiefern werden die einzelnen Mitarbeitenden je Ebene (Management) auf die Arbeit mit AI vorbereitet?
- Inwiefern wurde im Unternehmen ein spezielles Framing oder eine Storyline entwickelt?
- Inwiefern ist definiert, in welcher Organisationseinheit/-zelle welche Risk Awareness mit welchem Zeithorizont geschaffen werden muss, um strategische Vorteile zu nutzen, aber Risiken zu kontrollieren/mitigieren?

„Repository & Tools" umfasst die prozessuale Inventarisierung von AI-Systemen, die strukturierte Erfassung relevanter Systeminformationen und den Einsatz unterstützender Werkzeuge, um Dokumentation und Steuerung von Governance-Prozessen zu ermöglichen. Die erarbeiteten Fragen lauten:

- Inwiefern wird die Inventarisierung der AI-Systeme prozessual durchgeführt und mit welchen Mechanismen wird der Erfolg sichergestellt?
- Welche Informationen müssen je AI-System im AI-Repository enthalten sein?
- Inwiefern werden für das AI-Repository bzw. grundsätzlich für AI-Governance Tools eingesetzt und welche Funktionalitäten haben diese Tools?

Die Erkenntnisse werden in den Entwicklungspartnerschaften der CDR-Initiative unter Anwendung des CDR-Reifegradmodells kontinuierlich angewendet. Unterschiedliche Organisationsformen wie Krankenkassen, Versicherungsunternehmen, Mittelständler und Tech-Startups bringen fortlaufend ihre Perspektiven mit unterschiedlichsten Ausgangslagen ein. Gerade diese Vielfalt ist produktiv, denn sie führt zu erweiterten Kenntnissen über organisationsübergreifende AI-Adoption.

4. Praxisbeispiele gelingender KI-Adoption

4.1 Kultur & Literacy; Vom Datenverständnis zur Lernkultur

Das erste Praxisbeispiel beleuchtet, wie Kompetenzen im Umgang mit Daten über reines Verstehen hinausgehen und zur Entwicklung einer reflektierten, gemeinschaftlich getragenen Lernkultur beitragen. Dabei haben wir uns in den Gesprächen mit den Unternehmen an den im vorherigen Kapitel notierten Fragen orientiert.

Inwiefern werden die einzelnen Mitarbeitenden je Ebene (Management, …) auf die Arbeit mit AI vorbereitet?

Im Unternehmen wurde der Aufbau von AI-Kompetenz bewusst dezentral organisiert: Ein zentraler AI-Hub innerhalb der IT/Business Intelligence (BI) bündelt Methodik und technische Entwicklung, während in den Fachbereichen die operative Verantwortung für konkrete Use Cases liegt. Die Vorbereitung auf den Umgang mit AI erfolgt entlang klar definierter Rollen. Für Führungskräfte wurde ein aufwändiges, aber wirksames Schulungsmodul entwickelt, das in einem 8–10-stündigen Training zentrale Chancen und Herausforderungen des AI-Einsatzes adressiert und konkrete Handlungsoptionen mit Ansprechpartnern aufzeigt. Für Mitarbeitende wurden unter anderem ein Ambassadorenprogramm für Generative AI, interne Roadshows, offene AI-Sprechstunden und eine digitale Austauschplattform etabliert. Ein Schulungsnavigator ordnet Weiterbildungsangebote gezielt verschiedenen Zielgruppen zu – etwa Begeisterten, Neugierigen oder Zögernden. Grundlage bildet ein konzernweiter Data-Literacy-Pfad, der das Verständnis von Daten als Asset fördert und auch strategisch verankert ist. Ergänzt wird das Ganze durch ein internes, datenschutzkonformes GPT-System, das den begleiteten Praxiseinsatz erlaubt und ein gestuftes Verständnis für den produktiven, aber auch kritischen AI-Einsatz fördert.

Inwiefern wurde im Unternehmen ein spezielles Framing oder eine Storyline entwickelt?

Diese Maßnahmen sind eingebettet in eine klare kulturelle Erzählung: AI wird als notwendiger Bestandteil zukünftiger Wettbewerbsfähigkeit verstanden und damit als nächster Schritt in einem länger laufenden Kulturwandel, der Verantwortung und Nachhaltigkeit in den Mittelpunkt stellt. Die Storyline („AI verändert Arbeit, ersetzt sie aber nicht") dient als Orientierungsrahmen und wird gezielt über interne Formate kommuniziert. Ergänzend läuft ein konzernweites Projekt zur angestrebten AI-Kultur, in dem bestehende Lern- und Beteiligungsformate kartiert, Lücken identifiziert und ein Soll-Zustand für verschiedene Organisationseinheiten abgeleitet wird. Der Abgleich mit Benchmarks anderer Unternehmen unterstützt die strategische Verortung.

Inwiefern ist definiert, in welcher Organisationseinheit/-zelle, welche Risk Awareness mit welchem Zeithorizont geschaffen werden muss, um strategische Vorteile zu nutzen, aber Risiken zu kontrollieren/mitigieren?

Auch im Bereich Risk Awareness wird auf dezentrale Verantwortung gesetzt. Anstelle einer eigenen AI-Governance-Abteilung sind Aufgaben und Zuständigkeiten klar auf bestehende Linienfunktionen wie Datenschutz, Datensicherheit und Risikomanagement verteilt. Ziel ist eine Governance, die nicht nur technische Expertise, sondern auch Prozessverständnis und Risikokompetenz erfordert und diese gezielt fördert. Dabei wird allerdings deutlich, dass eine zu starke Fragmentierung Governance erschwert, während eine gewisse Zentralisierung Umsetzung erleichtert. Die Herausforderung liegt daher in der präzisen Klärung: Wer verantwortet was, wo, mit welchem Zeithorizont? Diese Fragen werden schrittweise bearbeitet: mit Fokus auf interdisziplinäre Zusammenarbeit, klaren Kontrollmechanismen und einer gemeinsamen Arbeitskultur, die AI als ein Mittel unter vielen versteht, um die strategischen Ziele des Unternehmens zu erreichen.

4.2 Repository & Tools: Vom Katalog zur Steuerungsfähigkeit

Das zweite Praxisbeispiel zeigt, wie Repositorien und Tools den Schritt vom reinen Katalogisieren hin zur aktiven Steuerungsfähigkeit ermöglichen. In den Gesprächen mit den Unternehmen haben wir uns dabei an den folgenden Fragen orientiert.

Inwiefern wird die Inventarisierung der AI-Systeme prozessual durchgeführt und mit welchen Mechanismen wird der Erfolg sichergestellt?

Im Unternehmen wurde der Aufbau eines AI-Repositories frühzeitig strategisch mit dem Ziel angegangen, nicht nur die Dokumentationspflichten zu erfüllen, sondern zugleich Steuerungswissen für Innovation und Risikomanagement zu schaffen. Der Prozess begann mit einer umfassenden Bestandsaufnahme in Tabellenkalkulationsprogrammen und wurde durch grundlegende Schulungen flankiert, die ein gemeinsames Begriffsverständnis („Was ist ein Modell? Was ist AI?") schufen. Darauf aufbauend wurde risikobasiert definiert, welche Systeme im Repository zu erfassen sind. Die Pflege erfolgt dynamisch: durch standardisierte, aufgabenbezogene Trainings, regelmäßige Aktualitätsabfragen und die Integration des Repositories in den Einführungsprozess neuer Produkte und Prozesse. Jedes neue System wird bereits in der Entwicklungsphase registriert, ein verpflichtendes Onboarding mit Risikoprüfung ist dabei vorgesehen. Die Verantwortung für die Pflege wird zunehmend den anwendungsfallbezogenen Ownern übertragen. Herausforderungen wie die Abgrenzung relevanter Systeme oder die Zusammenführung von Modell- und IT-Inventar werden pragmatisch und risikoorientiert gelöst.

Welche Informationen müssen je AI-System im AI-Repository enthalten sein?

Erfasst werden pro System unter anderem der zugehörige Anwendungsfall, die eingesetzte Technologie, der betroffene Geschäftsprozess, zentrale Datenquellen, verantwortliche Personen sowie eine generische Beschreibung mit Ausgangslage, Zielsetzung und Funktion des AI-Systems. Ergänzt wird dies durch Risikobetrachtungen (einschließlich Ausfallszenarien), eine Kostenanalyse, die Nutzenanalyse auf Basis einer Nutzwertanalyse sowie eine Kategorisierung der eingesetzten AI-Fähigkeiten nach unterschiedlichen Fähigkeiten wie Forecasting oder Anomalieerkennung. Perspektivisch sollen zusätzliche Informationen ergänzt werden, z. B. verwendete Software-Bibliotheken oder jeweils verfügbare Technologie-Expert:innen innerhalb des Unternehmens je nach AI-Technologie.

Inwiefern werden für das AI-Repository bzw. grundsätzlich für AI-Governance Tools eingesetzt und welche Funktionalitäten haben diese Tools?

Technisch wurde die Repository-Arbeit zunächst mit einer Sammlung in Tabellenkalkulationsprogrammen begonnen, mittlerweile aber in ein konfigurierbares Tool überführt, das eine strukturierte

Dokumentation des Inventars ermöglicht. Künftig sollen auch Freigabeprozesse und Workflows in das Tool integriert werden, um die Pflege zu erleichtern und durch verbindliche Prozesse abzusichern. Dabei gilt weiterhin das Prinzip: So wenig neue Tools wie möglich, so viele wie nötig: Es werden bestehende Systeme genutzt, wo sie Governance-Aufgaben bereits abdecken. Spezialsysteme kommen nur dort zum Einsatz, wo Governance- und Risikoanforderungen dies rechtfertigen. Voraussetzung für die Nutzung von Tools zur Erschließung von Automatisierungspotenzialen ist die erweiterte Kenntnis über Prozessschritte.

5. Erfolgsfaktoren von AI-Governance

Die Praxis zeigt, dass erfolgreiche AI-Adoption kein Zufallsprodukt ist, sondern das Ergebnis gezielter, aber pragmatischer organisatorischer Gestaltung. Aus den Expert*innengesprächen und Entwicklungspartnerschaften haben sich erste Erfolgsfaktoren herauskristallisiert, die in ihrer Kombination den Unterschied machen und die je nach Reifegrad einer Organisation unterschiedlich ausgestaltet werden müssen.

Ob in Versicherungen oder Tech-Unternehmen: die Basis erfolgreicher Governance ist es, einen Rahmen zu geben, ohne Innovation zu ersticken. Denn Governance ist kein Kontrollinstrument als Mittel zum Zweck. Erfolgreiche Organisationen haben dafür eigene Wege gefunden. Manche beginnen bewusst niedrigschwellig mit klaren Rollenbeschreibungen sowie einfachen, wenigen, aber verbindlichen Freigabeprozessen und so verankerten Prinzipien. Erfolgreich scheinen vor allem hybride Modelle aus zentraler Kompetenz- und Steuerungseinheit plus dezentraler und anforderungsgerechter Verantwortlichkeit in den Fachbereichen. Governance wirkt, wenn sie mitwächst. Nicht als allumfassende Architektur von Beginn an, sondern als iterative Struktur, die an echten Fällen lernt. Immer dann, wenn Governance anschlussfähig an bestehende Logiken ist und den Arbeitsalltag über individuelle Interessen hinaus unterstützt, entfaltet sie Wirkung.

Mit Blick auf die Learnings zu „Kultur & Literacy“ lässt sich Folgendes festhalten: AI-Kultur ist eher eine neue Form der gemeinsamen Arbeitskultur als eine Kultur mit AI im Mittelpunkt. Ein AI-System kann nicht erfolgreich ohne die Nutzer*innen gestaltet werden. Für eine erfolgreiche AI-Adoption ist zwar eine Entwicklung der Datenkompetenz bei den Führungskräften und

Mitarbeitenden Voraussetzung, etwa durch einen Datenkompetenzpfad zur Schaffung des Verständnisses von Daten als Asset. Aber AI-Literacy ist mehr als Wissen über Daten oder Algorithmen. Die technische Perspektive ist zwar notwendig, aber nicht hinreichend. Sie meint die Fähigkeit, AI in der eigenen Rolle zu verstehen und in dieser bewerten und gestalten zu können. Was muss wer wissen, können und entscheiden? Je nach Rolle (Führungskräfte, Fachbereiche etc.) und je nach Persona (Begeisterte, Neugierige, Ablehnende etc.) sollten daher Angebote geschaffen werden, um mit der Komplexität umzugehen, die ein disruptiver Technologiekomplex wie AI letztendlich darstellt. Organisationen, die hier effektiv sind, begreifen AI-Literacy weiterhin nicht als isolierte Schulungsoffensive, sondern als Resonanzraum als Zusammenspiel aus Kompetenzentwicklung, Haltung und Sinn. Es braucht Erfahrungsräume, in denen Unsicherheit zugelassen, Reflexion ermöglicht und konkrete Anwendung geübt werden können. Erfolgreiche Organisationen haben dafür unterschiedlichste Formate etabliert. Von Schulungsnavigatoren über Lernreisen, regelmäßigen interdisziplinären Sprechstunden, GenAI-Experimentierräumen bis hin zu organisationsweiten Austauschformaten, in denen AI nicht nur erklärt, sondern erlebt wird. Entscheidend ist dabei nie das Format allein, sondern dessen Einbettung in Alltag, Struktur und Strategie. Und dafür braucht es eine Erzählung, die AI nicht als Bedrohung, sondern als Gestaltungschance darstellt. In der Praxis bewährt haben sich etwa „AI verändert Arbeitsplätze, aber ersetzt sie nicht“, oder „Ohne AI würde es uns schon heute nicht mehr geben“. Unter Einbeziehung verschiedenster Organisationseinheiten wie Data Strategy, Controlling, Risikomanagement und auch der Unternehmensführung werden so unterschiedliche Perspektiven und Anliegen bei einem langjährigen Kulturwandel berücksichtigt.

Betrachten wir „Repository & Tools“ im Rückblick, so lassen sich ebenfalls wesentliche Erkenntnisse zusammenfassen: Gute AI-Governance beginnt nicht mit einem Tool, denn ein Tool ist nur so gut wie das Organisationsdesign, in das es eingebettet ist. Allerdings bleibt sie wirkungslos ohne Strukturen, die Verantwortung sichtbar und steuerbar machen. Die Anwendung von Repositories ist daher keine technische Spielerei, sondern das organisatorische Gedächtnis, in dem dokumentiert wird, was im Einsatz ist, wofür, mit welchen Zielen, Daten, Risiken etc. und vor allem mit welchen Zuständigkeiten und Verantwortungen. Einige Organisationen starten mit einfachen Excel-Listen oder Wikis. Der Grund ist kein Mangel, sondern Pragmatismus. Denn entscheidend ist nicht die Tooltiefe, sondern der Einbau in den Alltag:

Wer trägt Informationen ein? Wer prüft Aktualität? Wie wird Verantwortung im Prozess verankert? Mit der Beantwortung dieser Fragen entsteht Transparenz und damit Handlungsfähigkeit. Repositories entfalten Wirkung, wenn sie in der Governance-Logik mitgedacht werden, z. B. als Triggerpunkt im Risikomanagement. Fortgeschrittene Organisationen verknüpfen in Repositories Anwendungsfälle mit übergeordneten Wertschöpfungsprozessen und integrieren Literacy-Funktionen durch Anwendungsbeispiele oder Dos and Don'ts.

Für eine nachhaltige AI-Adoption sind sowohl kulturelle und kompetenzbezogene Voraussetzungen („AI-Kultur & Literacy") als auch organisatorische Strukturen und Instrumente („Repository & Tools") erforderlich. Die folgenden drei Empfehlungen haben sich aus unseren Erfahrungen mit den Unternehmen für eine gelingende AI-Adoption und damit eine Weiterentwicklung des Reifegrads bewährt.

Top-Erfolgsfaktoren „AI-Kultur & Literacy"

1. Gestalte AI als gemeinsame Arbeitskultur, die Verantwortung und Nachhaltigkeit als zentrale Werte umfasst, statt AI in den Mittelpunkt zu stellen!
2. Führe mit klaren Botschaften und schaffe Erfahrungsräume durch begleitetes „Learning by Doing", um Ängste abzubauen und AI in Richtung Chancen positionieren!
3. Befähige nach Rollen und Personas mit einem anforderungsgerechten Kompetenzmodell, mit besonderer Berücksichtigung von Führungskräften und Data Literacy als Fundament!

Top-Erfolgsfaktoren „Repository & Tools"

1. Nutze das Repository als organisatorisches Gedächtnis für organisatorisch sinnvolle (Steuerungs-)Funktionen wie Risikomanagement und nicht als Mittel zum Zweck für Compliance!
2. Starte pragmatisch mit einfachen Lösungen und kenne die Prozessschritte vor Automatisierung und Tooleinsatz!
3. Entwickle das Repository zu einer Innovations- und Lernplattform, um Anwendungsfälle zur Ideengenerierung und Bildung sichtbar zu machen!

6. Das Toolkit der CDR-Initiative für Erfolg in der Praxis

Gelingende AI-Adoption ist somit weder Produkt von Zufällen, einzelnen Persönlichkeiten oder dem Einsatz eines mächtigen Tools, noch ist es eine Blaupause, die sich beliebig auf andere Organisationen übertragen lässt. Die Erfahrungen durch die Projekte und Entwicklungspartnerschaften der CDR-Initiative zeigen: Was in einer Organisation funktioniert, wird nicht durch Nachahmung anschlussfähig, sondern durch Übersetzung in Strukturen, Rollen, Prozesse, Denkweisen. Es geht um die Fähigkeit, Verantwortung schrittweise und kontextsensibel durch klare Strukturen, differenzierte Literacy-Formate und eine Arbeitskultur aufzubauen, die AI nicht als weitere Technologieerscheinung, sondern als disruptiven Wandel begreift.

Das CDR-Reifegradmodell gibt Organisationen Orientierung und eine Sprache, um über Verantwortung im Digitalen auf einem Niveau zu sprechen, das zum eigenen Entwicklungspfad passt. Ob es um AI geht oder um andere Aspekte von Digitalverantwortung gibt das Modell eine klare Struktur zur Identifikation des nächsten sinnvollen Schritts. Und der CDR-Quick-Check dient dabei als Einstieg. Er gibt initial Hinweise darauf, wo die Organisation steht, und wo Risiken sowie Potenziale liegen. In den Entwicklungspartnerschaften wird er so zur Basis gemeinsamer Workshops, aus denen konkrete Prioritäten, Formate und Umsetzungslogiken entstehen, die auf Rolle, Struktur und Zielbild der jeweiligen Organisation zugeschnitten sind.

Literaturverzeichnis

Brink, A. (2022): Twin Transformation Management. Ethik als Schlüsselfaktor für die erfolgreiche Zwillingstransformation!, in: Herberger, T. (Hrsg.): Digitale Transformation und Nachhaltigkeit in der globalen Finanzwirtschaft, Baden-Baden: Nomos, 141–172.

Brink, A. / Esselmann, F. / Henke, L. (2025): Das CDR-Reifegradmodell. Rüstzeug für digitale Verantwortung in turbulenten Zeiten, in: Brink, A. (Hrsg.): Fairness in Zeiten Künstlicher Intelligenz, Baden-Baden: Nomos, 283–298.

CDR-Initiative (2025a): Corporate Digital Responsibility Initiative, URL: https://cdr-initiative.de/initiative (aufgerufen am 23/08/2025).

CDR-Initiative (2025b): Corporate Digital Responsibility-Kodex, URL: https://cdr-initiative.de/uploads/files/2021-06_Kodex_CDR-Initiative_2021-11-25-101231_cyzv.pdf (aufgerufen am: 23/08/2025).

Crosby, P. B. (2025): Quality is Free: The Art of Making Quality Certain, New York: McGraw-Hill.

Mäntymäki, M. / Minkkinen, M. / Birkstedt, T. / Viljanen, M. (2022). Putting AI ethics into practice: The hourglass model of organizational AI governance, DOI: 10.48550/arXiv.2206.00335.

Merbecks, U. (2023). Corporate Digital Responsibility (CDR) in Germany: Background and First Empirical Evidence From DAX 30 Companies in 2020, in: Journal of Business Economics, Jg. 94 / Nr. 7–8, 1025–1049, DOI: 10.1007/s11573-023-01148-6.

National Institute of Standards and Technology (2023): Artificial Intelligence Risk Management Framework (AI RMF 1.0), URL: https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf (aufgerufen am: 23/08/2025).

Nonaka, I. / Takeuchi, H. (1995): The Knowledge-Creating Company. How Japanese Companies Create the Dynamics of Innovation, Oxford: Oxford University Press.

Schein, E. H. / Schein, P. A. (2017): Organizational Culture and Leadership, Hoboken: John Wiley & Sons.

Autor*innen

Becker, Paula, ist Studierende des Masterprogramms „Philosophy & Economics“ an der Universität Bayreuth.

Bilfinger, Beatriz, ist Senior Programm Managerin für Digital Responsibility & Sustainability beim Bundesverband Digitale Wirtschaft (BVDW) e.V. Ihr Fokus liegt dabei auf verantwortungsvoller digitaler Transformation, moderner Arbeitskultur und der Umsetzung von Nachhaltigkeit in Unternehmen der digitalen Wirtschaft.

Blachetta, Marie Dr., ist Referentin für digitale Verantwortung bei der Initiative D21. In ihrer Tätigkeit geht sie insbesondere der Frage nach, wie wir digitale Zukünfte technologisch verantwortungsvoll gestalten können und welche Aspekte sozialer und ökologischer Nachhaltigkeit dabei eine Rolle spielen

Breuer, Emily, ist Absolventin des Bachelorstudienprogramms „Philosophy & Economics“ an der Universität Bayreuth und seit Herbst 2025 Studierende des Masterprogramms „Philosophy“ an der Universität Uppsala, Schweden

Brink, Alexander Prof. Dr. Dr., ist Universitätsprofessor für Wirtschafts- und Unternehmensethik an der Universität Bayreuth, Gründungspartner der CONCERN GmbH, Köln, und wissenschaftlicher Berater der CDR-Initiative, Berlin.

Carl, Valerie Dr., ist Postdoktorandin am Lehrstuhl für Wirtschaftsinformatik und Informationsmanagement an der Goethe Universität in Frankfurt und Unternehmerin. Sie befasst sich mit Forschung zu (unternehmerischer) digitaler Verantwortung, künstlicher Intelligenz (insbesondere der Erweiterung menschlicher Fähigkeiten durch KI), digitaler Resilienz und digitalen Ökosystemen.

Cerulli-Harms, Annette Dr., ist Projektmanagerin beim ConPolicy Institut für Verbraucherpolitik und seit 2025 stellvertretende Leiterin der Geschäftsstelle der CDR-Initiative. Zu ihren Themenschwerpunkten zählen die Bereiche Digitalisierung, Gaming, Bankenwesen und Finanzen, sowie Nachhaltigkeitsmanagement.

Christ, Anna-Sophie, ist Doktorandin an der Maastricht University und Universität Bremen. Im Bereich der Wirtschaftsinformatik forscht und lehrt sie zu digitaler Unternehmensverantwortung. In ihrer Arbeit als selbständiger Coach, Beraterin und System- und Organisationsauftstellerin begleitet sie Unternehmen zu Nachhaltigkeit und Digitalisierung.

Dähne, Jonah, ist Studierender des Bachelorprogramms „Philosophy & Economics“ an der Universität Bayreuth.

Dietl, Karin, ist Inhaberin der Beratungsagentur Value Design mit Fokus auf Digitalrecht und Data Compliance in Wien. 2021 initiierte sie die Business Plattform „CDR Austria" mit Veranstaltungen, einem Arbeitskreis und Ausbildungsangeboten rund um die CDR.

Dörr, Saskia Dr., ist Gründerin von WiseWay, Spezialistin für KI-Integration und Expertin für verantwortungsvolle Digitalisierung. Sie entwickelt KI-Agenten, die Unternehmen beim werteorientierten KI-Einsatz unterstützen.

Esselmann, Frank Dr., ist Partner bei der concern GmbH und berät zahlreiche Unternehmen sowie die CDR-Initiative des Bundesministeriums der Justiz und für Verbraucherschutz bei der digitalen Transformation. Seine Ausbildung und Forschungen in der Mathematik bilden den Hintergrund für seinen Fokus auf verantwortungsvolle, mehrwertstiftende Datenstrategien und Analysemethoden.

Förster, Maximilian, ist Bereichsleiter bei der kubus IT eGbR und Lehrbeauftragter für Digitalethik an der Hochschule Hof. Zudem forscht er an der Universität Bayreuth zum Thema Digitalverantwortung im Gesundheitswesen.

Hankins, Olivia, ist Studierende des Bachelorprogramms „Philosophy & Economics" an der Universität Bayreuth.

Hauer, Marc P., ist Senior AI Solution Architect am TÜV AI Lab. Darüber hinaus ist er als freiberuflicher Berater und Referent zum Themenkomplex KI an der Schnittstelle zur Gesellschaft aktiv, unter anderem im Auftrag der TrustedAI GmbH tätig.

Henke, Leonhard, ist als Projektmanager bei der concern GmbH in Köln tätig und berät die CDR-Initiative des Bundesministeriums der Justiz mit Fokus auf AI-Governance. Auf der Grundlage seiner interdisziplinären Expertise als Wirtschaftsingenieur legt er Wert auf pragmatische Ansätze, um technologische Innovationen mit nachhaltigem Geschäftserfolg in Einklang zu bringen.

Hinz, Maria, ist Teamleiterin für Gesunde Arbeit, Diversity, Nachhaltigkeit und Corporate Digital Responsibility (CDR) bei BARMER. Zuvor war sie als Digitalkoordinatorin für die unternehmensweite Verankerung von CDR verantwortlich. Sie bringt ihre Expertise in digitale Transformation und wertebasierte Unternehmensführung unter anderem in der CDR-Initiative des BMUV ein. Ihre beruflichen Stationen umfassen (Führungs-) Positionen in Politik, Verbänden und Unternehmen.

Hoffmann, Christian Hugo Dr., ist Philosoph und hat an der HSG in Finanzwissenschaft promoviert, er ist Gründer verschiedener Start-ups, Dozent an der Universität Zürich, Spezialist für Kryptowährungen, Technik¬ethiker, Liberaler und Experte für Künstliche Intelligenz.

Jänig, Jens-Rainer, ist Gründer und Inhaber der Markenagentur und Kommunikationsberatung mc-quadrat mit dem Themenschwerpunkt Transformationskommunikation. Von 2020

bis 2025 war er Herausgeber des ersten CDR-Online Magazins. 2014 gehörte er zu den initialen Initiatoren der CDR-Reflektion in Deutschland.

Keilhofer, Katharina, ist Referentin für Digitalisierung und Nachhaltigkeit für Verbraucherbelange am Bayerischen Staatsministerium für Umwelt und Verbraucherschutz (StMUV). Sie betreut die Initiative CDR Bayern sowie verschiedene Verbraucherbildungsangebote im Bereich Digitales.

Kettner, Sara Elisa Dr., ist Projektmanagerin beim ConPolicy Instituts für Verbraucherpolitik und seit 2023 Leiterin der Geschäftsstelle der CDR-Initiative. Zu ihren Themenschwerpunkten zählen verantwortungsvolle Digitalisierung und Nachhaltigkeit.

Klein, Till Dr., ist Head of AI Regulation bei appliedAI Institute for Europe GmbH. Er ist Mitglied des Netzwerks OECD AI und bei der Global Patnership of AI. Er arbeitet an Projekten, um die Konformität mit dem AI Act für Startups, KMU und dem öffentlichen Sektor zu beschleunigen.

Kurtz, Christian Dr., ist Postdoktorand am Arbeitsbereich IT-Management und Consulting der Universität Hamburg. Er forscht zu gesellschaftlichen Werten und Regulierungen in soziotechnischen Ökosystemen. Ein weiterer Schwerpunkt liegt dabei auch auf dem Verständnis des interdisziplinären Wirkens zwischen Wirtschaftsinformatik, Recht und Ethik, wobei Brückenkonzepte wie IT-Architekturen eine zentrale Rolle spielen.

Meier, Julius, ist Studierender des Bachelorprogramms „Philosophy & Economics" an der Universität Bayreuth.

Mihale-Wilson, Christina Dr., ist Forscherin und Entrepreneurin und arbeitet an der Schnittstelle zwischen Künstlicher Intelligenz, digitalen Ökosystemen und gesellschaftlicher Verantwortung.

Muff, Julia, ist Studierende des Bachelorprogramms „Philosophy & Economics" an der Universität Bayreuth.

Müller, Benjamin Prof. Dr., ist Professor für Digital Business an der Universität Bremen. Er forscht zur digitalen Transformation von Organisationen und zu verantwortungsvoller Digitalisierung.

Przybylek, Bartosz Alexander, ist Strategieberater mit dem Fokus auf Digitalisierung und Nachhaltigkeit. Bei ifok berät er Organisationen zu Corporate Digital Responsibility und unterstützt sie insbesondere bei der Strategieentwicklung und der Umsetzung. Zudem ist er ein gefragter Redner und Initiator von Dialogformaten zur digitalen Verantwortung.

Schaufler, Nicole, ist unabhängige Unternehmensberaterin und Coach für CDR & digitale Führungskultur, Teammitglied und Ambassador von CDR Austria und Mitglied des DIN-Arbeitsausschusses für Corporate Digital Responisibility. Ihre Arbeitsschwerpunkte liegen in der Informationssicherheit, KI-Compliance, Digitale Ethik und Green IT.

Schlegel, Isabelle Karoline, ist Managerin für KI und digitale Verantwortung bei Deloitte und Teil des Chief AI Office im Bereich Consulting. Als Mitbegründerin der „Digital Future Challenge" und Teilprojektleiterin des EU-Projekts „AI4Lungs" engagiert sie sich für interdisziplinäre Ansätze zur digitalen Transformation. Ihre Schwerpunkte liegen in der strategischen Kommunikation, digitalen Ethik und KI-Governance.

Scholz, Maike, ist Squad Lead Digital Ethics bei der Deutschen Telekom. Sie ist Vorstandsmitglied im Ressort CDR des BVDW und Referentin zu CDR, KI-Governance und ethischer Technologiegestaltung. Ihre Arbeitsschwerpunkte liegen in der Implementierung von Strategien zur verantwortungsvollen Nutzung von KI, der regulatorischen Umsetzung des EU AI Acts sowie der Förderung unternehmensweiter digitaler CDR-Standards.

Steinbeck, Kilian, ist Studierender des Bachelorprogramms „Philosophy & Economics" an der Universität Bayreuth.

Thorun, Christian Prof. Dr., ist Gründer und Geschäftsführer des ConPolicy Instituts für Verbraucherpolitik und Bereichsleiter Politik und Gesellschaft der Geschäftsstelle der CDR-Initiative. Er ist ein anerkannter Experte in der Vebraucherpolitik und berät Bundes- und Landesministerien sowie die Europäische Kommission in Verbraucherfragen.

Von Roesgen, Felix, ist Senior Consultant im Bereich Corporate Responsibility bei der ifok GmbH. Seine Themenschwerpunkte liegen an der Schnittstelle von Digitalisierung und Nachhaltigkeit, insbesondere den gesellschaftlichen Auswirkungen von Künstlicher Intelligenz.

Wagner, Julian, ist Studierender des Bachelorprogramms „Philosophy & Economics" an der Universität Bayreuth.

Wößner, Jakob, ist Vorsitzender im Ressort Digital Responsibility im Bundesverband der Digitalwirtschaft (BVDW e.V.). Er setzt sich für eine verantwortungsvolle Digitalisierung im Einklang mit Mensch und Natur ein. Sein Antrieb ist es, Daten nachhaltig zu nutzen und KI innovativ einzusetzen.