Fraunhofer IAO

Fraunhofer-Institut für Arbeitswirtschaft und Organisation IAO

Innovation im Blick

Vertrauenswürdige KI und verwandte Konzepte verstehen und anwenden

Ein Überblick für Unternehmen

Jessica Wulf | Diana Fischer-Pressler | Lydia Uhler | Marie Heidingsfelder | Simone Kaiser
Janika Kutz | Jens Neuhüttler

Im Auftrag des

IPAI

Inhalt

Executive Summary

Die Entwicklung und der Einsatz von vertrauenswürdigen KI-Systemen eröffnen die Möglichkeit, KI nachhaltig und ethisch verantwortungsvoll zu nutzen. Dadurch werden Risiken minimiert und Potenziale gehoben. Aber was ist vertrauenswürdige KI eigentlich und wie steht sie im Zusammenhang mit Konzepten wie erklärbare KI, ethische KI, faire KI, grüne KI, menschenzentrierte KI und verantwortliche KI? Antworten auf diese Fragen sowie Empfehlungen für Unternehmen vertrauenswürdige KI zu gestalten und zu nutzen, gibt dieses Whitepaper.

Dafür wird das Konzept der vertrauenswürdigen KI der HLEG on AI (2019) als Ausgangspunkt gewählt, da es auf einem holistischen Ansatz basiert und wissenschaftlich sowie politisch relevant ist. Die Konzepte erklärbare KI, faire KI, menschenzentrierte KI, grüne KI, ethische KI und verantwortliche KI sind in der wissenschaftlichen Diskussion eng mit vertrauenswürdiger KI verbunden und werden in diesem Whitepaper mit den Anforderungen an vertrauenswürdige KI verglichen und analysiert. Dabei werden Überschneidungen und Differenzen in Bezug auf vertrauenswürdige KI herausgearbeitet.

Vertrauenswürdige KI umfasst ethische Leitlinien, die sicherstellen sollen, dass KI-Systeme rechtmäßig, ethisch und robust sind. Dies bedeutet, dass sie gesetzlichen Anforderungen entsprechen, ethischen Prinzipien folgen und sowohl technisch als auch sozial robust sind. Daraus ergeben sich sieben Hauptanforderungen an vertrauenswürdige KI-Systeme: **1** Vorrang menschlichen Handelns und menschliche Aufsicht, **2** Technische Robustheit und Sicherheit, **3** Datenschutz, Datenqualitätsmanagement und Privatsphäre **4** Transparenz, **5** Vielfalt, Nicht-Diskriminierung und Fairness, **6** Gesellschaftliches und ökologisches Wohlbefinden sowie **7** Rechenschaftspflicht.

Die Analyse ergibt, dass die Konzepte erklärbare KI, ethische KI, faire KI, grüne KI, menschenzentrierte KI und verantwortliche KI sich in ihren Anforderungen stark mit den Anforderungen an vertrauenswürdige KI-Systeme überschneiden. Der Unterschied liegt in den Schwerpunkten der jeweiligen Konzepte.

Trotz ihrer Relevanz und der starken Überschneidungen werden Konzepte, wie vertrauenswürdige KI, heute wenig in der unternehmerischen Praxis umgesetzt. In diesem Whitepaper werden folgende Herausforderungen identifiziert und Handlungsempfehlungen für Unternehmen abgeleitet:

1. Konzepte wie vertrauenswürdige KI erheben den Anspruch der Allgemeingültigkeit – sie sollen für sehr unterschiedliche KI-Systeme sowie unterschiedliche Anwendungskontexte gelten. Dementsprechend sind die Konzepte abstrakt und wenig handlungsleitend. Vertrauenswürdige KI soll vor diesem Hintergrund als Grundgerüst verstanden werden, das für den jeweiligen konkreten Anwendungsfall im Unternehmen spezifisch operationalisiert werden sollte.

2. Der technische Diskurs zu KI-Systemen verläuft derzeit meist getrennt von ethischen Diskursen. Für eine effektive Umsetzung von Konzepten wie vertrauenswürdiger KI sind interdisziplinäre Teams in der Entwicklung und Implementierung von KI-Systemen wichtig, in denen Technik und Ethik von Anfang an zusammengedacht werden.

3. Anforderungen an vertrauenswürdige KI stehen im Spannungsverhältnis zueinander und teilweise zur Leistungsfähigkeit der Systeme. Deswegen sollte im konkreten Anwendungsfall unternehmensindividuell zwischen den jeweiligen Anforderungen und in Bezug auf Leistungsfähigkeit abgewogen und das spezifische Risiko bedacht werden.

4. Vermenschlichung von KI-Systemen führt zu Verantwortungsdiffusion. Deswegen braucht es eindeutig geklärte Verantwortlichkeiten innerhalb des Unternehmens in allen Entwicklungs- und Anwendungsschritten von KI-Systemen.

5. Partizipation fehlt im Konzept der vertrauenswürdigen KI und sollte bei der Einführung von KI-Systemen in Unternehmen mitgedacht werden: Raum für Mitbestimmung und Bedenken der Mitarbeitenden sollte sichergestellt werden.

KI-Systeme verantwortungsvoll zu entwickeln und einzusetzen ist notwendig, um die Potenziale von KI-Systemen im Unternehmenskontext nutzbar zu machen. Unternehmen sollten sich daher aktiv mit dem Thema auseinandersetzen und Leitlinien konsequent für sich operationalisieren und umsetzen. Dies bietet nicht nur die Möglichkeit, Potenziale von KI-Systemen voll auszuschöpfen, sondern stärkt auch das nachhaltige Vertrauen in die Technologie. Mit dem richtigen Ansatz können Unternehmen eine Vorreiterrolle in der Nutzung vertrauenswürdiger KI übernehmen und gleichzeitig ethische sowie gesellschaftliche Standards voranbringen.

1. Einleitung

Die rasante Entwicklung von künstlicher Intelligenz (KI) bringt vielfältige Potenziale und Möglichkeiten mit sich: Diese reichen von der Optimierung von Prozessen über die Steigerung von Effizienz und Qualität bis hin zu Produktivitätszuwächsen, die beispielsweise durch schnellere und umfangreichere Datenverarbeitung sowie der Entwicklung neuer Services erzielt werden können. Doch zeigt sich immer wieder, dass KI-Systeme auch signifikante Risiken mit sich bringen. So können Entscheidungsprozesse, die durch KI-Systeme automatisiert werden, zu intransparenten Ergebnissen führen. Zudem besteht die Gefahr der Verstärkung bestehender Vorurteile und Diskriminierungen, da Vorhersagemodelle von KI-Systemen auf bereits vorhandenen Daten basieren, die verzerrt sein können (für Beispiele siehe Seite 7). Eine zentrale Bedeutung erhält in diesem Zusammenhang das Konzept der vertrauenswürdigen KI, das ethische Leitlinien beinhaltet, die diese Risiken reduzieren und verhindern sollen. Dadurch soll der positive Einsatz von KI-Systemen und die effektive Integration in diverse Anwendungsbereiche unterstützt werden.

In Reaktion auf die dringende Notwendigkeit, Risiken von KI-Anwendungen zu minimieren, haben Unternehmen, akademische Einrichtungen, Regierungen und Interessensgruppen in den vergangenen Jahren zahlreiche ethische Leitlinien mit dem Ziel herausgegeben, ethische Anforderung an KI-Systeme zu definieren. Aus unterschiedlichen fachlichen und regulatorischen Perspektiven ist inzwischen eine komplexe und schwer überschaubare Vielfalt von Leitlinien, Konzepten und Ansätzen entstanden.[1] Darüber hinaus wird die Diskussion um vertrauenswürdige KI durch die Vielzahl an Begrifflichkeiten, die im Zusammenhang mit vertrauenswürdiger KI stehen, zusätzlich komplex. Konzepte, die hierbei immer wieder auftauchen, sind beispielsweise »faire KI«, »erklärbare KI« oder »ethische KI«.

Aus Sicht von Unternehmen kann die Vielfalt der Leitlinien und Konzepte eine Herausforderung darstellen und zu Fragen führen wie: »Welche Leitlinien sind für uns relevant?« und »Welche Konzepte müssen wir im Rahmen unserer Entwicklungsarbeit berücksichtigen?«. Die Vielfalt an Begrifflichkeiten kann Verwirrung stiften und die Entwicklung klarer und kohärenter betrieblicher Strategien zur Implementierung und Steuerung von KI-Technologien erschweren. Aus diesem Grund greift das Whitepaper die primär diskutierten Konzepte rund um vertrauenswürdige KI auf und nimmt eine Differenzierung und Definition der Begriffe vor. Ziel ist es Unternehmen eine fundierte und praktikable Auseinandersetzung mit dem Thema vertrauenswürdiger KI und verwandter Konzepte zu ermöglichen und eine Orientierung für die Adaption in der unternehmerischen Praxis zu geben.

Ausgangspunkt der Analyse ist das Konzept der vertrauenswürdigen KI und dessen Kernanforderungen der Europäischen Kommission - High Level Expert Group on AI (HLEG). Ausgehend von diesem Konzept wurde eine Analyse der Fachliteratur durchgeführt, um die Konzepte der erklärbaren, fairen, menschenzentrierten, grünen, ethischen und verantwortungsvollen KI gegenüberzustellen. Dabei wurden Unterschiede und Gemeinsamkeiten zwischen den Konzepten und ihren Kernanforderungen herausgearbeitet. Darüber hinaus werden Herausforderungen bei der Umsetzung von diesen Kernanforderungen beschrieben und Handlungsempfehlungen für Unternehmen abgeleitet.

1 Zu den zahlreichen Unternehmen, die ethische Leitlinien für KI-Systeme entwickelt haben, gehören bei-spielsweise die Deutsche Telekom, SAP, Sony, Telefonica, Deep Mind, Accenture, Microsoft, IBM, OpenAI oder Google. Auf internationaler Policy-Ebene haben die OECD, die Europäische Kommission, UNESCO, die IEEE und die G20 Leitlinien veröffentlicht. Aus der Zivilgesellschaft kommen die ethischen Leitlinien bei-spielsweise von Amnesty International, Partnership on AI, Access Now, UNI Global Union oder privacy in-ternational.

Beispiele für Risiken von KI-Systemen

Bewerbungssoftware von Amazon
Die Bewerbungssoftware von Amazon sollte Bewerbungen vorsortieren und geeignete Bewerberinnen und Bewerber vorschlagen. Noch in den Testphasen stellte sich heraus, dass die Bewerbungen von Frauen systematisch aussortiert, also nicht empfohlen wurden. Der zugrunde liegende Algorithmus wurde mit den Einstellungsdaten der letzten zehn Jahre von Amazon trainiert. Selbst nach technischen Anpassungen, bei denen die Variable »Geschlecht« aus dem Modell entfernt wurde, blieb es bei der systematischen Benachteiligung von weiblichen Bewerberinnen. Dieses Beispiel illustriert deutlich, wie Biases und ungerechte Entscheidungen, die in historischen Daten verborgen sind, durch KI-Systeme nicht nur reproduziert, sondern möglicherweise verstärkt werden können (Dastin 2018).

Autonomes Fahren
Während der Paralympics in Tokyo 2020 wurde ein Athlet von einem selbstfahrenden Auto angefahren und konnte daher nicht am Wettkampf teilnehmen. Toyota stellte autonom fahrende Shuttle-Services im Athleten-Dorf zur Verfügung. Der sehbeeinträchtigte Athlet Aramitsu Kitazono überquerte gerade eine Kreuzung, als er von dem Shuttle angefahren wurde. Da der Athlet in Richtung des Busses schaute, wurde angenommen, dass er den Bus bemerkt hatte und anhalten würde. Dies war jedoch nicht der Fall. Als Reaktion darauf wurden die autonomfahrenden Busse im Olympischen Dorf eingestellt (Shivdas und Kelly 2021).

Qualitätskontrolle von Airbags
Ein weiteres hypothetisches Beispiel ist der Einsatz von KI bei der Qualitätsprüfung von Airbags im Automobilbereich. Wurde die KI beispielsweise primär mit Daten trainiert, die bestimmte Fehlertypen nicht abbilden, könnte sie fehlerhafte Airbags übersehen. Auch nach technischen Anpassungen, wie der Integration zusätzlicher Sensordaten, könnten kritische Defekte unerkannt bleiben. Dies hätte zur Folge, dass Airbags, die nicht den Sicherheitsstandards entsprechen, dennoch in Fahrzeuge eingebaut werden und damit das Risiko für die Insassen potenziell erhöhen. Dieses hypothetische Beispiel verdeutlicht die kritische Bedeutung der Qualität und Vollständigkeit der Daten, die für das Training von KI-Systemen in sicherheitsrelevanten Anwendungen verwendet werden.

2. Vertrauenswürdige KI

Angesichts der inhärenten Risiken, die mit dem Einsatz von KI-Systemen einhergehen, haben zahlreiche Länder, Organisationen und Unternehmen ethische Anforderungen an KI in speziellen Rahmenwerken und Leitlinien formuliert (Jobin et al. 2019). Diese Initiativen gehen mit der Einschätzung der Europäischen Kommission (European Commission, 2021) einher: Menschen können die Vorteile einer Technologie nur dann vollständig nutzen, wenn sie dieser vertrauen. Die Sicherstellung der Vertrauenswürdigkeit von KI-Systemen wird daher als wesentlich für die Förderung ihrer Entwicklung und Implementierung betrachtet (Thiebes et al. 2020). Vor diesem Hintergrund definieren Kaur et al. (2022) vertrauenswürdige KI folgendermaßen:

»Vertrauenswürdige KI ist ein Orientierungsrahmen, mit dem sichergestellt wird, dass ein System auf der Grundlage von Beweisen für seine angegebenen Anforderungen vertrauenswürdig ist. Es stellt sicher, dass die Erwartungen der Nutzerinnen und Nutzer und der Interessengruppen auf nachprüfbare Weise erfüllt werden«

(Kaur et al. 2022, S.4)

Die Vertrauenswürdigkeit von Technologien, einschließlich KI, wird durch die vertrauensgebende Person anhand verschiedener Dimensionen bewertet. Darunter fallen auch solche, die typischerweise Menschen zugesprochen werden. Beispielsweise stellten Benbasat und Wang (2005) fest, dass Nutzende von Online-Produktempfehlungsagenten, diesen Systemen menschliche Attribute wie Wohlwollen und Integrität zuschreiben. Dies verdeutlicht, dass das Vertrauen in Technologie auf ähnlichen Grundlagen wie das zwischenmenschliche Vertrauen basieren kann. Gleichzeitig spielen auch systembezogene Vertrauensdimensionen eine wichtige Rolle wie bspw. Funktionalität, Nützlichkeit und Zuverlässigkeit (Mcknight et al. 2011), welche in Tabelle 1 definiert werden.

Tabelle 1: Beispiele von Dimensionen wahrgenommener systembezogener Vertrauenswürdigkeit

Dimensionen	Definition
Funktionalität	Überzeugung, dass die betreffende Technologie durchweg ordnungsgemäß funktionieren wird.
Nützlichkeit	Überzeugung, dass die betreffende Technologie angemessene und reaktionsschnelle Hilfe für die Nutzer/-innen bietet.
Zuverlässigkeit	Überzeugung, dass die betreffende Technologie verlässlich funktionieren wird.

Vertrauen in KI-Systeme hängt damit nicht nur von der messbaren, technologischen Leistungsfähigkeit ab, sondern auch von der subjektiven Beurteilung ihrer ordnungsgemäß und angemessen funktionierenden Systemeigenschaften durch deren Nutzerinnen und Nutzer. Selbst bei einer Einstufung eines Systems nach bestimmten Leitlinien als vertrauenswürdig, kann nicht automatisch von einem Entstehen von Vertrauen ausgegangen werden (Pieters 2011; Reinhardt 2023). Dennoch sind Leitlinien wichtig, da sie grundlegende Anforderungen formulieren, um Vertrauenswürdigkeit zu ermöglichen und so als Orientierung dienen.

Vertrauen und KI

Die Vertrauensforschung hat ihren Ursprung im Vertrauen in zwischenmenschlichen Beziehungen. In diesem Zusammenhang wird Vertrauen als ein psychologischer Zustand verstanden, in dem Verletzlichkeit auf der Grundlage positiver Erwartungen in Bezug auf die Absichten oder das Verhalten einer anderen Person akzeptiert wird (Lukyanenko et al. 2022). Vertrauen ist kein eindimensionales Konzept, sondern ein vielschichtiges Konstrukt, das je nach Kontext und den an einer Beziehung beteiligten Akteuren variieren kann.

Das Konzept des Vertrauens wird inzwischen auch auf technische Systeme wie KI-basierte Systeme angewendet (Ameen et al. 2021). Im Kontext von KI ist Vertrauen ein wesentlicher Faktor für deren Akzeptanz, Fortschritt und Entwicklung (Thiebes et al. 2020). Das Vertrauen in KI-basierte Systeme wird von verschiedenen Faktoren beeinflusst, darunter die Funktionalität der Technologie, ihre Interpretierbarkeit, Zuverlässigkeit und Interaktionsschnittstellen. Darüber hinaus wird das Vertrauen in KI-Systeme auch durch den Kontext beeinflusst, in dem sie eingesetzt werden, z. B. durch das spezifische Anwendungsgebiet oder die Branche.

Ein entscheidender Faktor im Vertrauensprozess ist die Inkaufnahme von Unsicherheit, die zum einen aus der vom Menschen wahrgenommenen eigenen Verletzlichkeit resultiert. Zum anderen aus dem Risiko der Unklarheit über die Absichten von Personen oder die Funktionsweise bzw. die Ziele, für die KI-Systeme eingesetzt werden (Lee und See 2004). Die Rolle der Unsicherheit ist bei der Interaktion mit KI-Systemen relevant, insbesondere bei Verfahren des Deep Learnings mit neuronalen Netzen, bei denen Transparenz und Nachvollziehbarkeit über Entscheidungswege und Funktionsweisen der KI-Systeme fehlen. Wie aber wird entschieden, ob einer Person oder einem System vertraut wird oder nicht? Hier ist die Vertrauenswürdigkeit der Person oder des Systems von Bedeutung.

3. Grundsätze und verwandte Begriffe von vertrauenswürdiger KI

In diesem Kapitel werden im ersten Teil die Grundsätze dargestellt, die die Vertrauenswürdigkeit von KI-Systemen sicherstellen sollen. Als Ausgangspunkt und Referenzrahmen dieser Darstellung dient die Definition und Beschreibung vertrauenswürdiger KI-Systeme, die die HLEG on AI 2019 im Auftrag der Europäischen Kommission erarbeitet und veröffentlicht hat.

Die High Level Expert Group on AI
Die Europäische Kommission hat im Jahr 2018 eine Gruppe von Sachverständigen ernannt, die sie bei der Strategie für KI beraten und unterstützen soll: die High-Level Expert Group on Artificial Intelligence (EC HLEG on AI 2019). Die EK HLEG setzte sich aus 52 Expertinnen und Experten zusammen, darunter 24 aus der Wirtschaft, 17 aus der Wissenschaft, 5 aus der Zivilgesellschaft sowie 6 weiteren Personen. Neben den »Ethik-Leitlinien für eine vertrauenswürdige KI« veröffentlichte die EK HLEG auch Empfehlungen für politische Maßnahmen und Investitionen in Bezug auf KI-Systeme.

Der Ansatz der EK HLEG dient aus drei Gründen als Grundlagenpapier: (1) Zum einen wird der Veröffentlichung der EK HLEG die Popularisierung des Konzepts zugeschrieben, der ein signifikanter Anstieg an wissenschaftlichen, politischen und unternehmerischen Papieren zu dem Konzept vertrauenswürdiger KI folgten (Freiman 2023). (2) Zum anderen dient dieses Papier als Grundlage und Vorüberlegung für den vor kurzem verabschiedeten EU AI-Act, der die Gesetzgebung in Bezug auf KI-Systeme in Deutschland prägt und beeinflusst. (3) Darüber hinaus ist der Ansatz, der EK HLEG ganzheitlich in Bezug auf die unterschiedlichen Phasen in der Entwicklung und Anwendung von KI-Systemen. Das Konzept vertrauenswürdiger KI bezieht sich auf den gesamten Zyklus, also formuliert Anforderungen an die Entwicklung, Implementierung und Nutzung der Systeme (Kaur et al. 2022).

Der EU AI Act
Der EU AI Act (deutsch: KI-Verordnung) ist die weltweit erste umfassende Regulierung für KI-Systeme. Laut Europäischem Parlament ist das Ziel des AI-Acts, »dass KI-Systeme, die in der EU verwendet werden, sicher, transparent, nachvollziehbar, nicht-diskriminierend und umweltfreundlich sind« (European Parliament 2023). Der EU AI-Act basiert auf dem Konzept der vertrauenswürdigen KI der HLEG.

Im zweiten Teil dieses Kapitels wird das Konzept vertrauenswürdiger KI im Zusammenhang mit verwandten Begriffen und Konzepten analysiert. Das Feld Entwicklung und Anwendung von KI-Systemen vertrauenswürdig zu gestalten, wirkt unübersichtlich und ist durchaus komplex: Neben vertrauenswürdiger KI werden auch Begriffe wie erklärbare KI, verantwortliche oder ethische KI verwendet. Darüber hinaus stecken hinter diesen Begriffen teilweise die gleichen und teilweise unterschiedliche Anforderungen an KI-Systeme. Die Komplexität lässt sich zum einen durch den Versuch erklären, dass diese ethischen Leitlinien für alle Arten von KI-Systemen und alle Branchen gültig sein sollen. Zum anderen beschäftigen sich viele unterschiedliche Akteurinnen und Akteure aus unterschiedlichen Disziplinen und Traditionen mit dem Thema und bringen dementsprechend unterschiedliche Perspektiven und Ansätze mit. Im folgenden Kapitel werden die Begriffe eingeordnet.

3.1 Grundsätze und Anforderungen an vertrauenswürdige KI

In ihrem Bericht »Ethik-Leitlinien für eine vertrauenswürdige KI« (EK HLEG on AI 2019) definiert die EK HLEG vertrauenswürdige KI als rechtmäßig, ethisch und robust. Damit ein KI-System als vertrauenswürdig gelten kann, muss es über den gesamten Lebenszyklus des Systems alle anwendbaren Gesetze und Bestimmungen einhalten (rechtmäßig), ethischen Grundsätzen und Werten entsprechen (ethisch) und sowohl technisch als auch sozial robust sein (robust). Technische und soziale Robustheit bedeutet, dass das KI-System technisch einwandfrei funktionieren muss und darüber hinaus auch in sozialen sich verändernden Anwendungskontexten zuverlässig funktioniert. Rechtmäßigkeit, Ethik und Robustheit müssen zusammengedacht werden und gleichermaßen garantiert sein, damit ein System als vertrauenswürdig gelten kann.

Aus diesen drei Komponenten ergeben sich sieben wesentliche Anforderungen an die Entwicklung, Einführung und Nutzung von KI-Systemen, die erfüllt werden müssen, damit ein KI-System als vertrauenswürdig gilt. In Tabelle 2 werden die sieben Anforderungen an vertrauenswürdige KI zusammengefasst.

Die EK HLEG betont, dass diese sieben Hauptanforderungen an KI-Systeme gemeinsam und nicht isoliert betrachtet werden müssen. Sie gelten von der Entwicklung des KI-Systems über die Implementierung sowie die gesamte Nutzdauer des KI-Systems. Sie beziehen sich zum einen auf technische Komponenten, wie die Funktionsweise der KI-Modelle. Zum anderen auf soziale Komponenten, wie die Akteurinnen und Akteure sowie Prozesse, zum Beispiel durch das Einbeziehen diverser Stakeholder in den Entwicklungsprozess der KI-Systeme oder die Offenlegung von Geschäftsmodellen.

Tabelle 2: Anforderungen an vertrauenswürdige KI basierend auf den »Ethik-Leitlinien für eine vertrauenswürdige KI« der High Level Expert Group on AI der Europäischen Kommission (2019)

Anforderung	Erläuterung
Vorrang menschlichen Handelns und menschliche Aufsicht	KI-Systeme unterstützen Menschen, ohne sie zu ersetzen; Erhaltung menschlicher Autonomie durch Gewährleistung der menschlichen Aufsicht; Wahrung der Grundrechte.
Technische Robustheit[1] und Sicherheit	Erbringung der von Nutzenden erwarteten Leistung; bei Fehlern eine Minimierung von Schäden, Zuverlässigkeit durch Sensibilität gegenüber Inputänderungen, Widerstandsfähigkeit gegen Angriffe; Reproduzierbarkeit der Ergebnisse.
Datenschutz und Datenqualitätsmanagement, Privatsphäre	Schutz der Privatsphäre und Daten der Nutzenden, Qualität und Integrität der Daten im gesamten Lebenszyklus der KI, Schutz vor Missbrauch, klare Zugriffsregeln und fristgerechte Datenvernichtung.
Transparenz	Rückverfolgbarkeit und Erklärbarkeit von Entscheidungsprozessen durch Offenlegung der Daten, Systeme und Geschäftsmodelle; Nutzende und von den Systemen betroffene Personen verstehen die Leistungen und Grenzen des Systems.
Vielfalt, Nicht-Diskriminierung und Fairness	Vermeidung von direkter und indirekter Diskriminierung sozialer Gruppen; Berücksichtigung und Einbindung aller Stakeholder in den gesamten Lebenszyklus.
Gesellschaftliches und ökologisches Wohlergehen	Verhinderung von gesellschaftlichen Schäden oder Umweltschäden in der Entwicklung, Implementierung und Nutzung; Nachhaltigkeit und positive soziale Veränderungen durch Berücksichtigung von Umweltschutz und sozialen Auswirkungen.
Rechenschaftspflicht	Nachvollziehbarkeit und Rechtfertigung der Entscheidungen basierend auf dem KI-System; klare Verantwortlichkeiten für alle richtigen und falschen Entscheidungen, Überprüfbarkeit der Ergebnisse.

1 In der Definition von vertrauenswürdiger KI ist mit Robustheit soziale und technische Robustheit gemeint. In den konkreten Anforderungen ist unter der Anforderung Robustheit lediglich die technische Robustheit gemeint. Die Aspekte, die unter soziale Robustheit fallen, wurden in anderen Anforderungen berücksichtigt.

3.2 Verwandte Konzepte und Zusammenhänge

Neben vertrauenswürdiger KI gibt es weitere verwandte Konzepte: erklärbare KI, faire KI, menschenzentrierte KI, grüne KI, ethische KI und verantwortliche KI. Diese Konzepte sind in der wissenschaftlichen Diskussion eng mit dem Konzept der vertrauenswürdigen KI verbunden – teilweise werden sie synonym verwendet. Generell zielen alle diese Konzepte darauf ab, die Technologie sicherer, transparenter und nutzerfreundlicher zu gestalten, haben jedoch einen unterschiedlichen Fokus. Auf Seite 12 und 13 sind diese Konzepte alphabetisch aufgelistet und definiert.

Die Analyse der Konzepte basiert auf Metastudien zu den jeweiligen Begriffen: erklärbare KI, faire KI, ethische KI, menschenzentrierte KI, grüne KI und verantwortliche KI. Diese Metastudien wurden anhand von drei Kriterien ausgewählt: (1) Aktualität der Studien (Veröffentlichung nach 2019), (2) wissenschaftliche Standards (Studien haben Peer-Review-Verfahren durchlaufen) und (3) Relevanz in der wissenschaftlichen Community (Suche nach viel zitierten Studien). Im folgenden wurden die in den Metastudien erarbeiteten Definitionen zusammengefasst.

Begriff **Erklärbare KI**

Erklärung

»Erklärbare KI (Explainable AI, XAI) zielt darauf ab, eine Reihe von Techniken des maschinellen Lernens bereitzustellen, die es menschlichen Nutzenden ermöglichen, Modelle zu verstehen, ihnen angemessen zu vertrauen und sie erklärbarer zu machen«.

(Dwivedi et al. 2023, S. 194)

Der Ansatz der erklärbaren KI ist eine Antwort auf das sogenannte Black-Box-Problem: Bei vielen KI-Modellen ist nicht nachvollziehbar, wie die Ergebnisse zustande kommen (Angelov et al. 2021). Ansätze der erklärbaren KI zielen darauf ab, die den Modellen zugrundeliegenden Entscheidungswege für Menschen verständlich zu machen. Erklärbarkeit gilt als Voraussetzung für die Entscheidung sich auf die Empfehlung des KI-Systems zu verlassen, oder nicht. Erklärbare KI ist notwendig für die gesellschaftliche Akzeptanz von KI-Systemen und Voraussetzung für eine sinnvolle Regulierung (Angelov et al. 2021).

Begriff **Ethische KI / KI-Ethik**

Erklärung

»Im Zusammenhang mit der KI beschreibt die Ethik der KI die moralischen Verpflichtungen und Pflichten einer KI und ihrer Entwickelnden«.

(Siau und Wang 2020, S. 75)

In diesem Zusammenhang wird zwischen KI-Ethik und ethischer KI unterschieden. Die KI-Ethik befasst sich mit ethischen Grundsätzen und Regeln und wendet diese auf KI-Systeme an. KI-Ethik steht eng im Zusammenhang mit menschlichen ethischen Fragen, zum Beispiel Fragen rund um Gerechtigkeit, Privatsphäre oder Transparenz.

Ethische KI bezeichnet ein KI-System, dem ethisch korrektes Verhalten zugeschrieben wird und das festgelegten moralischen Verpflichtungen entspricht, sowohl in der Entwicklung als auch in der Anwendung des KI-Systems.

Faire KI

»Der Begriff *faire KI* bezieht sich auf probabilistische Entscheidungshilfen, die eine ungleiche Schädigung (oder einen ungleichen Nutzen) verschiedener Untergruppen verhindern (Barocas und Selbst 2016). Bei fairer KI besteht das Ziel darin, Systeme bereitzustellen, die sowohl Vorurteile quantifizieren als auch die Diskriminierung von Untergruppen abschwächen«.

(Feuerriegel et al. 2020, S. 379)

Das Konzept der fairen KI bezieht sich im Besonderen auf KI-Systeme, die Entscheidungsprozesse durch Vorhersagen unterstützen; beispielsweise durch das Vorsortieren vielversprechender Lebensläufe aus einem Pool von Lebensläufen. Faire KI zielt darauf ab, individuelle Fairness sicherzustellen, indem Personen unabhängig von ihrer Zugehörigkeit zu sozialen Gruppen, bei gleichen Voraussetzungen gleichbehandelt werden (individuelle Fairness). Zusätzlich wird Gruppen-Fairness angestrebt, damit insbesondere marginalisierte oder historisch diskriminierte Gruppen nicht benachteiligt werden (Gruppen-Level Fairness).

Menschenzentrierte KI

»Menschenzentrierte KI nutzt Daten, um ihre menschlichen Nutzenden zu unterstützen und zu befähigen, während es gleichzeitig die zugrundeliegenden Werte, Vorurteile, Grenzen und die Ethik der eigenen Datenerfassung und Algorithmen offenlegt, um eine ethische, interaktive und anfechtbare Nutzung zu fördern«.

(Capel und Brereton 2023, S. 13)

Das Konzept der menschenzentrierten KI legt einen klaren Fokus auf die Auswirkungen, die die Anwendung von KI-Systemen auf den Menschen haben kann. Deswegen sollen die Entwicklung, das Design, sowie die Implementierung darauf ausgelegt sein, Nutzende zu unterstützen und zu ermächtigen. Zentral ist dabei, dass der Anwendungskontext von Beginn an berücksichtigt wird, da er maßgeblich die Auswirkungen bestimmt.

Grüne KI

»Grüne KI bezieht sich auf Praktiken, die darauf abzielen, KI zu nutzen, um die Auswirkungen des Menschen auf die natürliche Umwelt in Bezug auf die genutzten natürlichen Ressourcen und/oder die Auswirkungen, die KI selbst auf die natürliche Umwelt haben kann, zu mindern«.

(Verdecchia et al. 2023, S. 17)

Grüne KI ist der Oberbegriff für zwei Konzepte, die den Zusammenhang zwischen KI-Systemen und Umwelt beschreiben. Das Konzept *Grün in KI* befasst sich mit den direkten und indirekten negativen Auswirkungen von KI-Systemen auf die Umwelt, die durch die Entwicklung, dem Einsatz oder der Nutzung sowie dem Ende der Funktionsdauer der Systeme ergeben (Weber et al. 2023). Darunter fallen beispielsweise Ansätze, die den hohen Energieverbrauch beim Training von KI-Systemen reduzieren sollen.

Das Konzept *Grün durch KI* beschäftigt sich mit KI-Systemen, die mit dem Ziel entwickelt und eingesetzt werden, Prozesse nachhaltiger zu gestalten und zu optimieren (Weber et al. 2023). Darunter fallen beispielsweise Systeme, die durch eine gezieltere Taktung Stromverbrauch reduzieren sollen.

Verantwortliche KI

Das Konzept der verantwortlichen KI ist eine »Methodik für die groß angelegte Implementierung von KI-Methoden in realen Organisationen, bei der Fairness, Erklärbarkeit der Modelle und Verantwortlichkeit im Mittelpunkt stehen«.

(Barredo Arrieta et al. 2020, S. 82)

Verantwortliche KI wird als ein Konzept beschrieben, das unterschiedliche ethische Prinzipien und Anforderungen vereint und dadurch die verantwortliche Nutzung von KI-Systemen in Organisationen ermöglicht. Die Basis des Konzepts ist erklärbare KI, die eine Voraussetzung für die ethischen Anforderungen auf dem Weg zu einer verantwortungsvollen Nutzung von KI ist.

Es wird deutlich, dass es bezüglich der Anforderungen an vertrauenswürdige, erklärbare, ethische, faire, grüne, menschenzentrierte und verantwortliche KI-Systeme einige Schnittstellen gibt.

Um die Schnittstellen zwischen den Konzepten zu verdeutlichen, werden die Anforderungen von erklärbarer KI, ethischer KI, fairer KI, grüner KI, menschenzentrierter KI, und verantwortlicher KI in Abbildung 1 ins Verhältnis mit den sieben Hauptanforderungen an vertrauenswürdige KI der EK HLEG gesetzt. Für den Abgleich wurde für jedes Konzept eine Metastudie identifiziert und nach den drei oben genannten Kriterien ausgewählt. Um die Metastudien zu ergänzen und Zusammenhänge besser analysieren zu können, wurden ergänzende wissenschaftliche Artikel einbezogen.

Die Relevanz der Anforderungen an vertrauenswürdige KI für die jeweiligen Konzepte wurde auf Basis der gewählten Metastudien für Abbildung 1 klassifiziert. Da sich die Metastudien sowohl methodisch als auch konzeptionell unterscheiden, musste je nach Metastudie bei der Klassifizierung unterschiedlich vorgegangen werden. Sofern eine Gewichtung und Priorisierung der Anforderungen durch das Autorenteam der Metastudie direkt vorgenommen wurde, wurde diese in die Abbildung 1 übertragen. War das nicht der Fall, wurde eine qualitative Einschätzung auf Basis der Ergebnisse der Metastudie und der Expertise der Autorinnen dieses Whitepapers vorgenommen.

Die Größe der Punkte richtet sich danach, wie prominent eine Anforderung in der Studie genannt wurde oder wie direkt sie im Zusammenhang mit dem jeweiligen Konzept diskutiert wurde. Wird die Anforderung an vertrauenswürdige KI in der Beschreibung des Konzepts in der Metastudie nicht explizit adressiert, findet sich kein Punkt an der Stelle. Abbildung 1 soll als Orientierungshilfe dienen und einen Überblick über die vielfältigen Konzepte und Anforderungen in diesem komplexen und unübersichtlichen Feld geben.

Das Konzept der fairen KI definiert beispielsweise die Anforderung *Vielfalt, Nicht-Diskriminierung und Fairness* als wichtigste Anforderung für das Konzept. *Rechenschaftspflicht* wird auch als Anforderung formuliert, jedoch als nachgelagert bzw. als Ergänzung. Die anderen fünf Referenzanforderungen werden bei fairer KI nicht direkt adressiert und haben deswegen keinen Punkt in der Abbildung.

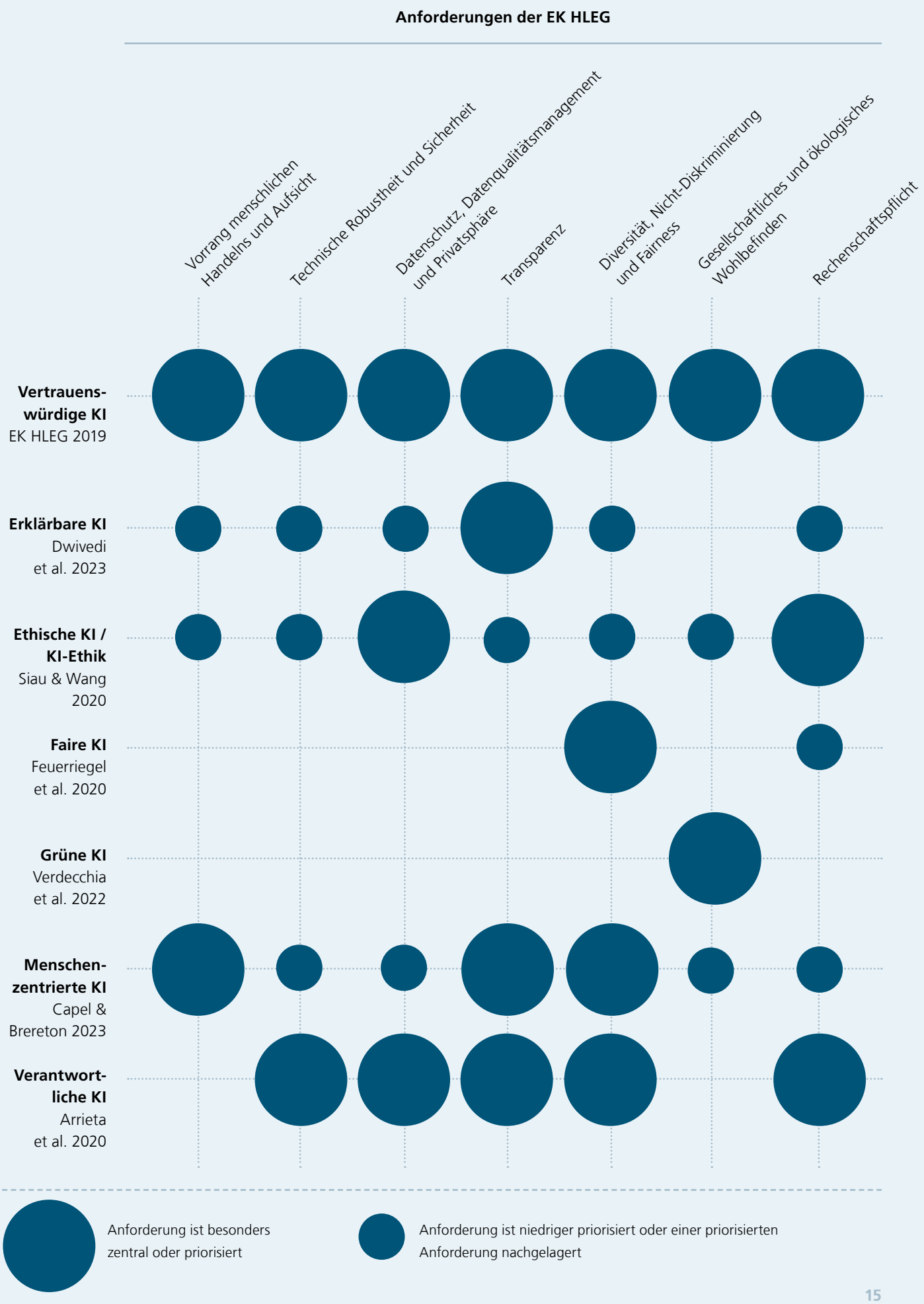
Der Überblick in Abbildung 1 zeigt, dass sich mindestens eine, meist jedoch mehrere Anforderungen der EK HLEG an vertrauenswürdige KI in den untersuchten Metastudien zu den Konzepten finden. Schnittstellen ergeben sich aus Abhängigkeiten unter den Anforderungen sowie zwischen den Konzepten. Unterschiede entstehen aus der Gewichtung und Priorisierung der Anforderung in den Konzepten. Im Folgenden werden diese Unterschiede und Gemeinsamkeiten für jedes Konzept genauer beschrieben.

Erklärbare KI

Durch erklärbare KI sollen Menschen, die den Modellen zugrunde liegenden Entscheidungswege nachvollziehen können. Laut der Metastudie von Dwivedi et al. (2023) ist erklärbare KI eng mit der Anforderung *Transparenz* verbunden. Unter dem Begriff »erklärbare KI« werden konkrete technische Methoden zusammengefasst, die eine Datenerklärbarkeit und Modellerklärbarkeit gewährleisten sollen. Dadurch wird die Grundlage für Transparenz geschaffen. Der Datensatz muss jedoch offengelegt werden, damit die Methoden angewendet werden können. Darüber hinaus müssen die verwendeten Modelle erklärbar sein, da nur so eine Transparenz gewährleistet werden kann. In diesem Kontext werden die Begriffe *Erklärbarkeit* und *Transparenz* teilweise synonym verwendet.

Die Anforderung *Transparenz* beinhaltet in der eigenen Definition die Erklärbarkeit von Entscheidungsprozessen und stellt somit die Hauptanforderung dar, die für erklärbare KI erfüllt sein muss.

Abbildung 1: Überblick der Anforderungen ähnlicher Konzepte im Vergleich mit den Hauptanforderungen an vertrauenswürdige KI der EK HLEG.


Anforderungen der EK HLEG
Vorrang menschlichen Handelns und Aufsicht
Technische Robustheit und Sicherheit
Datenschutz, Datenqualitätsmanagement und Privatsphäre
Transparenz
Diversität, Nicht-Diskriminierung und Fairness
Gesellschaftliches und ökologisches Wohlbefinden
Rechenschaftspflicht
Vertrauens-würdige KI EK HLEG 2019
Erklärbare KI Dwivedi et al. 2023
Ethische KI / KI-Ethik Siau & Wang 2020
Faire KI Feuerriegel et al. 2020
Grüne KI Verdecchia et al. 2022
Menschen-zentrierte KI Capel & Brereton 2023
Verantwort-liche KI Arrieta et al. 2020
Anforderung ist besonders zentral oder priorisiert
Anforderung ist niedriger priorisiert oder einer priorisierten Anforderung nachgelagert

Die Anforderungen *Vorrang menschlichen Handelns, Robustheit und Sicherheit*, *Datenqualitätsmanagement, Vielfalt, Nicht-Diskriminierung und Fairness sowie Rechenschaftspflicht* sind nachgelagerte Anforderungen: Sie können erst erfüllt werden, wenn Erklärbarkeit und Transparenz gegeben sind. Demnach ist Transparenz grundlegende Voraussetzung dafür, dass die anderen Anforderungen umgesetzt werden können (Dwivedi et al. 2023). Beispielsweise kann die Anforderung Rechenschaftspflicht nur dann erfüllt werden, wenn Transparenz gegeben ist und das Modell erklärbar ist. Rechenschaftspflicht gewährleistet, dass Entscheidungen basierend auf KI-Systemen nachvollziehbar und gerechtfertigt sind. Rechenschaftspflicht umfasst auch klare Verantwortlichkeiten für eventuelle negative Auswirkungen von KI-Systemen. Für transparente Entscheidungen braucht es eine Offenlegung der Entscheidungsprozesse und der Funktionsweise des KI-Systems, was wiederum die Notwendigkeit von Erklärbarkeit und Transparenz als Grundvoraussetzung für Rechenschaftspflicht unterstreicht.

Das Konzept der erklärbaren KI umfasst zudem die Erklärbarkeit der dem System zu Grunde liegenden Daten. Die Offenlegung der verwendeten Daten ist eine Voraussetzung für die Überprüfung der Qualität und Integrität dieser Daten. Dementsprechend ist die Anforderung *Datenqualitätsmanagement* auf die Erklärbarkeit des Datensatzes angewiesen. Auch die Anforderung *menschliches Handeln und Kontrolle* ist von der Erklärbarkeit der Daten und Modelle abhängig. Nur wenn diese Daten und Modelle für Menschen nachvollziehbar sind, können sie Empfehlungen des KI-Systems effektiv überwachen und bewerten (Dwivedi et al. 2023; Jobin et al. 2019; Chamola et al. 2023). Das bedeutet auch, dass Konzepte, die die Anforderung *Vorrang menschlichen Handelns und Kontrolle* priorisieren stark von der Transparenz und Erklärbarkeit des KI-Systems abhängig sind. Auch technische Robustheit wird durch Transparenz unterstützt und verbessert (Chamola et al. 2023). Modelle, die transparent und erklärbar sind, sind tendenziell auch robuster. Der Grund dafür ist, dass eine kohärente Erklärung für Vorhersagen bedeutet, dass die Argumentation logisch ist. Eine logische Argumentation wiederum ist weniger anfällig für Störungen.

Faire KI

Wie die Bezeichnung faire KI nahelegt, fokussiert das Konzept die Anforderung *Vielfalt, Nicht-Diskriminierung und Fairness* (Feuerriegel et al. 2020). Zudem ist für faire KI die Anforderung der *Rechenschaftspflicht* entscheidend, da sie klare Verantwortlichkeit ermöglicht. Diese Klarheit über Verantwortlichkeiten hilft dabei, faire Verhaltensweisen zu fördern und rechtliche Unsicherheiten zu minimieren, falls ein KI-System die Anforderung *Vielfalt, Nicht-Diskriminierung und Fairness* nicht erfüllt. Wenn die Anforderung der *Rechenschaftspflicht* erfüllt ist, ist unter anderem eindeutig geregelt, wer verantwortlich ist, sollte es beispielsweise zu Diskriminierung durch ein KI-System kommen. Dementsprechend unterstützt Rechenschaftsplicht auch die Anforderung nach Fairness (EC HLEG on AI 2019).

Ethische KI

Ethische KI wird unter Berücksichtigung der Anforderungen der KI-Ethik entwickelt. Die KI-Ethik adressiert alle sieben der Hauptanforderungen von vertrauenswürdiger KI, wobei Datenschutz, Datenqualitätsmanagement und Privatsphäre sowie Rechenschaftspflicht in der Metastudie hervorgehoben werden. Unterschiede bestehen in der Benennung der Anforderungen. In der Metastudie zu ethischer KI werden einige der Anforderungen an ethische KI durch Fragen definiert, die nicht beantwortet werden. Beispielsweise die Anforderung »ethische Standards«: »Ohne umfassende und unvoreingenommene ethische Standards, wie können Menschen eine Maschine dazu trainieren, ethisch zu handeln?« (Siau und Wang 2020, S. 18)

Für ethische KI gelten darüber hinaus die Anforderungen *Demokratie und Bürgerrechte« sowie »Automatisierung und Wegfall von Arbeitsplätzen*. Diese Anforderungen werden bei vertrauenswürdiger KI in Teilen über die Anforderung *gesellschaftliches und ökologisches Wohlergehen* thematisiert. Bei ethischer KI werden beide Anforderungen über Negativbeispiele definiert. Unethische KI führt demnach dazu, dass Demokratien geschwächt werden und Bürgerrechte nicht mehr für alle Bürger und Bürgerinnen gelten. Automatisierung und Wegfall von Arbeitsplätzen definiert keine konkreten Anforderungen an KI-Systeme, sondern reflektiert die Diskussion, welche potenziellen Auswirkungen KI-Systeme auf den Arbeitsmarkt haben könnten.

An ethische KI wird darüber hinaus die Anforderung *Menschenrechte* gestellt. In diesem Kontext bedeutet das, dass Entwickler und Entwicklerinnen die Menschenrechte kennen müssen, um sie nicht aus Versehen zu missachten (Siau und Wang 2020). Im Konzept der vertrauenswürdigen KI sind die Menschenrechte nicht als Anforderung an die Systeme definiert, sondern Rahmen, in dem das Konzept der vertrauenswürdigen KI verankert ist (EC HLEG on AI 2019).

Grüne KI
Bei grüner KI liegt der Fokus auf dem Zusammenhang zwischen KI-Systemen und ökologischen Auswirkungen. Demnach ist hier die Anforderung gesellschaftliches und ökologisches Wohlbefinden zentral. Dabei sollen KI-Systeme für eine bessere Ressourcennutzung und zur Reduktion des ökologischen Fußabdruckes eingesetzt werden (grün durch KI). Der Fokus der Metastudie liegt jedoch auf *Grün in KI,* damit ist das Bestreben gemeint, KI-Technologie an sich umweltfreundlicher und energieeffizienter zu gestalten. Ethik spielt in diesem Konzept nur nachgelagert eine Rolle: In Bezug auf die ethischen Auswirkungen, die steigende CO2-Fußabdrücke von KI-Systemen haben.

Theoretisch bezieht sich das Konzept auf alle Phasen im Lebenszyklus von KI-Systemen, wie Design, Training, Einsatz und Betrieb. Tatsächlich liegt der Fokus jedoch meist auf der Trainingsphase von Algorithmen des maschinellen Lernens und vernachlässigt die restlichen Phasen (Verdecchia et al. 2023). Für die Trainingsphase werden konkrete technische Methoden vorgeschlagen, wie die ressourcenschonende Gestaltung von Algorithmen. Darüber hinaus werden Modelle vorgeschlagen, die den Punkt der optimalen Genauigkeit der Modelle bei einem geringstmöglichen Energieverbrauch angeben (Verdecchia et al. 2023).

Grüne KI und Vertrauenswürdigkeit
Das Konzept der grünen KI mag auf den ersten Blick weit von dem Konzept der vertrauenswürdigen KI entfernt sein. Doch basierend auf dem Konzept der EK HLEG on AI ist »Nachhaltigkeit und ökologische Verantwortung« (EK HLEG on AI 2019, S. 23) eine essentielle Anforderung an eine vertrauenswürdige KI, weil sie auf die Grundsätze Fairness und Schadensverhütung einzahlen – neben Menschen müssen auch »andere fühlende Wesen und die Umwelt als Akteur berücksichtigt werden« (EC HLEG on AI 2019, S. 23). Dabei geht es der EK HLEG on AI (2019) ähnlich wie in dem Konzept der grünen KI um den Verbrauch von Ressourcen und Energie sowie die Umweltverträglichkeit der Systeme.

Menschenzentrierte KI
Menschenzentrierte KI umfasst alle sieben der Hauptanforderungen an vertrauenswürdige KI sowie die Konzepte der *ethischen und erklärbaren KI* (Capel und Brereton 2023). Das Konzept der *menschenzentrierten KI* zielt darauf ab den Menschen und seine Bedürfnisse in das Zentrum von KI-Systemen zu stellen. Daraus ergeben sich Anforderungen, die über die Anforderungen der EK HLEG an vertrauenswürdige KI hinausgehen: Menschenzentrierte KI legt einen Fokus auf Mensch-Maschine-Interaktion und ein menschenzentriertes Design, beispielsweise bei Benutzeroberflächen (Capel und Brereton 2023). Menschliche Anforderungen an KI-Systeme werden über technische Anforderungen gestellt. Der Fokus auf den Menschen in der Entwicklung und Anwendung von KI-Systemen soll zu einer faireren, gerechteren und nachhaltigeren Gesellschaft beitragen (Ozmen Garibay et al. 2023).

Verantwortliche KI
Verantwortliche KI beinhaltet die Konzepte der erklärbaren KI und ethische Leitlinien – also KI-Ethik. Laut Definition von verantwortlicher KI muss diese erklärbar und fair sein – das Konzept der erklärbaren KI und fairen KI ist also in das Konzept der verantwortlichen KI integriert (Arrieta et al. 2020). Erklärbare KI ist in diesem Konzept ähnlich wie oben beschrieben die Voraussetzung dafür, dass die Anforderungen *Fairness, Transparenz und Privatsphäre* erfüllt und überprüft werden können (Arrieta et al. 2020). Dementsprechend überschneiden sich die Anforderungen an verantwortliche KI mit fünf der sieben Anforderungen an vertrauenswürdige KI und sind alle gleich gewichtet: *Technische Robustheit und Sicherheit, Datenschutz, Datenqualitätsmanagement und Privatsphäre, Transparenz, Vielfalt, Nicht-Diskriminierung und Fairness* und zuletzt auch *Rechenschaftspflicht*.

Zusammenfassend zeigt die Analyse auf, dass die Konzepte erklärbare KI, ethische KI, faire KI, grüne KI, menschenzentrierte KI und verantwortliche KI viele Schnittstellen mit den sieben Hauptanforderungen an vertrauenswürdige KI aufweisen. Jedes dieser Konzepte bringt spezifische Schwerpunkte ein, die abhängig vom Anwendungsbereich und den Zielen eines KI-Systems variieren können. Die Erfüllung der Anforderungen in den jeweiligen Konzepten tragen wesentlich dazu bei, KI-Systeme verantwortungsvoll zu gestalten und ihre Akzeptanz zu erhöhen. Dies ist die Grundlage für den erfolgreichen Einsatz von KI-Systemen in vielfältigen Lebens- und Wirtschaftsbereichen.

Die Vielzahl an Konzepten und deren Überschneidungen zeigt jedoch auch die Komplexität des Themenfeldes auf: alle Technologien, die als KI verstanden werden sowie die Anwendung in unterschiedlichen Kontexten allgemeingültig verantwortungsvoll und ethisch zu gestalten.

4. Herausforderungen und Empfehlungen für die Umsetzung vertrauenswürdiger KI

Die Hauptanforderungen des Konzepts *vertrauenswürdiger KI* der EK HLEG (1) Vorrang menschlichen Handelns und menschliche Aufsicht, (2) Technische Robustheit und Sicherheit, (3) Datenschutz, Datenqualitätsmanagement und Privatsphäre (4) Transparenz, (5) Vielfalt, Nicht-Diskriminierung und Fairness, (6) Gesellschaftliches und ökologisches Wohlbefinden sowie (7) Rechenschaftspflicht erweisen sich – im Abgleich mit den parallel entstandenen anderen Konzepten wie ethische oder verantwortliche KI – als gute und umfassende Grundlage, um sich mit ethischen Anforderungen an KI-Systeme zu beschäftigen.

Trotz ihrer Relevanz und der starken Überschneidungen haben die analysierten Konzepte (Kapitel 3) heute noch wenig Einfluss auf die unternehmerische Praxis (Hagendorff 2020). Im Folgenden werden die verschiedenen Gründe dafür dargestellt und passende Handlungsempfehlungen abgeleitet, um die Anforderungen an vertrauenswürdige KI in Unternehmen zu implementieren:

Konzept der vertrauenswürdigen KI als Orientierungshilfe und Einstieg in das Thema

Ethische Leitlinien, wie die für vertrauenswürdige KI, erheben den Anspruch für alle KI-Systeme gültig zu sein, die in unterschiedlichen Branchen und Kontexten angewendet werden. Die Leitlinien gelten für Empfehlungssoftware im Onlinehandel, die für gezielte Werbung (Targeted Advertisement) genutzt werden ebenso wie für Systeme, die Lebensläufe in Bewerbungsverfahren vorsortieren. Chatbots, die Nutzende unterstützen sollen sich auf Webseiten oder bei Unternehmensangeboten zurecht zu finden, werden ebenso adressiert, wie Systeme des autonomen Fahrens. Die Auswirkungen können variieren, von der Empfehlung ungeeigneter Services an Kundinnen und Kunden bis hin zu potenziell lebensbedrohlichen Situationen, wie dem Überfahren von Menschen. Diese breite Anwendbarkeit führt dazu, dass die Leitlinien auf einem hohen Abstraktionsniveau formuliert sind und wenig spezifische Anweisungen bieten. Sie decken ein Spektrum an Anwendungsfällen ab, die jeweils eigene Risiken und mögliche schwere Konsequenzen bergen, sollte die Technologie fehlerhaft sein.

Um auf alle technischen Systeme und Anwendungskontexte zu passen, müssen ethische Leitlinien, die beanspruchen allgemeingültig zu sein, auf einem hohen Level und wenig spezifisch bleiben. Das bedeutet gleichzeitig – und das ist einer der Hauptkritikpunkte an dem Konzept der *vertrauenswürdigen KI* und ähnlichen Konzepten – dass es keine präzisen und eindeutigen Entscheidungshilfen für die Entwicklung und Anwendung von KI-Systemen vorgibt (Stahl und Leach 2023). Ähnlich wie Leitlinien der OECD, NATO und WHO (vgl. Schmitt 2022) veröffentlicht auch die Europäische Kommission hauptsächlich »Richtlinien« oder »Prinzipien« für vertrauenswürdige KI, die in der praktischen Umsetzung nicht konkret genug sind, um effektive Handlungsanweisungen zu bieten (Brundage et al. 2020; Mittelstadt 2019).

Die Allgemeingültigkeit ethischer Leitlinien führt zu Unklarheiten bei ihrer Operationalisierung im Unternehmenskontext, da konkrete Bewertungskriterien fehlen (Stahl und Leach 2023; Schmitt 2022). Ein typisches Beispiel ist die Anforderung der Transparenz, die sich in vielen Leitlinien wieder findet. Wie Transparenz in der Praxis umgesetzt wird und woran konkret festgemacht wird, ob die Anforderung erfüllt wurde, wird entweder gar nicht adressiert, oder unterscheidet sich je nach Definition (Jobin et al. 2019). Unterschiedliche Interpretationen darüber, was Transparenz konkret bedeutet – ob es um die Offenlegung des Quellcodes, die Einschränkungen des Systems oder die potenziellen Auswirkungen geht – führen zu Inkonsistenzen in den Vorgaben und deren Anwendung (Jobin et al. 2019).

Handlungsempfehlung 1

In der Praxis bedeutet das für Unternehmen, dass die Leitlinien für vertrauenswürdige KI ein Grundgerüst und Ausgangspunkt bieten. Das Konzept der *vertrauenswürdigen KI* liefert Orientierung, um konkrete Regeln und Handlungsschritte für den eigenen spezifischen Anwendungsfall abzuleiten. Dabei müssen Fragen beantwortet werden, was für ein KI-System für welches Problem im Unternehmen eingesetzt werden soll oder was spezifisch für den Anwendungskontext (Branche, Anwendung) beachtet werden muss. Für diesen konkreten Anwendungskontext können die Anforderungen für vertrauenswürdige KI dann durch das Unternehmen spezifisch und individuell operationalisiert und messbar gemachen werden.

Die Entwicklung von Design Prinzipien können hier hilfreich sein, um die Leitlinien in praktische Lösungen umzusetzen. Sie dienen als Brücke zwischen den theoretischen Anforderungen und der praktischen Umsetzung. Basierend auf den allgemeinen Leitlinien müssen Unternehmen spezifische Anforderungen für ihren Kontext ableiten. Das bedeutet, dass sie analysieren müssen, welche Aspekte der Leitlinien für ihren spezifischen Anwendungsfall relevant sind. Diese Prinzipien müssen dann in den Entwicklungsprozess integriert und regelmäßig überprüft und angepasst werden, um sicherzustellen, dass das KI-System nicht nur funktional und effizient, sondern auch vertrauenswürdig ist.

Zusammenführen ethischer Anforderungen und technischer Entwicklungen durch Interdisziplinarität

Ebenfalls herausfordernd für die praktische Umsetzung der Anforderungen für vertrauenswürdige KI und ähnliche Konzepte ist, dass die Diskurse über ethische Anforderungen und technische Entwicklungen von KI-Systemen meist getrennt geführt und nicht zusammengedacht werden (Hagendorff 2020).

Handlungsempfehlung 2

Unternehmen, die KI-Systeme entwickeln, sollten auf interdisziplinär zusammengesetzte Entwicklungsteams achten, sodass ethische Anforderungen von Anfang an bedacht und KI-Systeme entsprechend entwickelt werden. Unternehmen, die KI-Systeme einsetzen wollen, sollten in der Entscheidung für ein spezifisches System sowie in der Implementierung ebenfalls auf Interdisziplinarität achten und so ethische Anforderungen einfließen lassen.

Abwägung zwischen Anforderungen und Leistungsfähigkeit des KI-Systems

Eine weitere Herausforderung in Bezug auf die praktische Umsetzung der Anforderungen an *vertrauenswürdige KI* besteht darin, dass diese Anforderungen teilweise im Konflikt mit der Leistungsfähigkeit der Systeme stehen (Thiebes et al. 2020). Das ist zum Beispiel der Fall bei Transparenz und Erklärbarkeit von KI-Modellen, da erklärbare Modelle oft weniger leistungsfähig sind (Knight 2017).

Darüber hinaus widersprechen sich die Anforderungen teilweise: Die Erfüllung einer Anforderung erschwert oder verhindert die Erfüllung einer anderen Anforderung. Um der Anforderung *Vielfalt, Nicht-Diskriminierung und Fairness* nachzukommen, sollen immer größere und vor allem repräsentativere Datensätze generiert werden. Die Sammlung von mehr Daten, einschließlich sensibler personenbezogener Daten, steht im Konflikt mit dem Erfordernis der Privatsphäre und dem datenschutzrechtlichen Grundsatz der Datensparsamkeit (Jobin et al. 2019; Thiebes et al. 2020). Des Weiteren entsteht dabei auch eine Spannung zu der Anforderung des gesellschaftlichen und ökologischen Wohlergehens: Die Speicherung und Verarbeitung großer Datensätze erhöht den Energieverbrauch und den Wasserverbrauch für die Kühlung (Weidinger et al. 2022; Gillin 2021).

Handlungsempfehlung 3

Je nach KI-System und Anwendungskontext muss einerseits zwischen den sieben Anforderungen abgewogen und priorisiert werden. Andererseits muss zwischen den sieben Anforderungen und der Leistungsfähigkeit des KI-Systems priorisiert werden. Dabei sind je nach Anwendungskontext unterschiedliche Risiken zu berücksichtigen und abzuwägen. So ist es z. B. bei Entscheidungen, die tief in die Integrität von Menschen eingreifen oder denen sich Menschen nicht entziehen können, wie z. B. bei Bewerbungsverfahren oder Bonitätsprüfungen, unerlässlich, dass im Zweifelsfall nachvollzogen werden kann, wie die Entscheidung zustande gekommen ist und dass die Entscheidung erklärt werden kann.

Vermeidung von Verantwortungsdiffusion und Vermenschlichung

Neben diesen allgemeinen Kritikpunkten an ethischen Leitlinien für KI-Systeme, existieren auch spezifische Kritikpunkte am Konzept der *vertrauenswürdigen KI*. Wie in Kapitel 2 dargestellt kommt viel der Forschung zu Vertrauen aus zwischenmenschlichen Beziehungen und die Erkenntnisse werden auf Vertrauen in KI-Systeme übertragen. Dabei ist es wichtig, dass es nicht zu einer Vermenschlichung von KI-Systemen kommt, die mit einer Verantwortungsdiffusion einhergehen: Für Risiken und negative Auswirkungen von KI-Systemen braucht es klare Verantwortlichkeiten von Menschen und Unternehmen. Verantwortung kann nicht an die KI-Systeme übergeben werden (Freiman 2023).

Handlungsempfehlung 4

Die Festlegung klarer Verantwortlichkeiten im Entwicklungsprozess von KI-Systemen ist essenziell, um vertrauenswürdige KI zu gewährleisten: Durch die klare Zuweisung von Zuständigkeiten für jede Phase – von der Konzeption über die Entwicklung bis hin zum Einsatz – wird sichergestellt, dass ethische Standards eingehalten, Qualitätskontrollen durchgeführt und Risiken effektiv entgegengewirkt werden.

Handlungsempfehlung 5

Wenn ein Unternehmen KI-Systeme einsetzen möchte, ist es unerlässlich, dass die Verantwortlichkeiten zwischen dem anwendenden Unternehmen und dem Unternehmen, das das System entwickelt, klar geregelt sind. Dazu ist es notwendig, dass umfassende Informationen über das System für die Anwendung zur Verfügung gestellt werden. Nur so können Risiken angemessen eingeschätzt und eine fundierte Verantwortung für den Einsatz des KI-Systems übernommen werden.

Berücksichtigung von Akzeptanz und Partizipation der Mitarbeitenden

Ein weiterer Kritikpunkt an dem Konzept der *vertrauenswürdigen KI* ist, dass Partizipation zu Vertrauen beiträgt. Dieser Aspekt ist in den Anforderungen jedoch nicht angemessen berücksichtigt. Die Anforderungen Akzeptanz und Partizipation von Individuen und Gruppen fehlen in vielen Leitlinien (Kaur et al. 2022). Dementsprechend ist es wichtig, vor allem Menschen, die von den Auswirkungen von KI-Systemen betroffen sind und solche, die Bedenken haben, in die Entwicklung und Einführung von KI-Systemen einzubeziehen.

Handlungsempfehlung 6

Partizipation und Akzeptanz der Mitarbeitenden sollten von Anfang an mitgedacht werden: Bereits bei der Planung der Einführung von KI-Systemen im Unternehmen sollten Momente der Mitbestimmung und Partizipation der Mitarbeitenden eingeplant werden. Bei der Umsetzung ist es wichtig, auch kritische Fragen und Perspektiven ernst zu nehmen, auf Ängste einzugehen und Mitgestaltung zu ermöglichen.

Die Vorgaben ethischer Richtlinien, zum Beispiel für die Schaffung einer vertrauenswürdigen KI, sind zusammenfassend häufig durch eine hohe Allgemeinheit und eine geringe Präzision charakterisiert. Das bedeutet für die praktische Anwendung in Unternehmen, dass diese Anforderungen individuell interpretiert und auf den spezifischen Anwendungskontext zugeschnitten werden müssen. Innerhalb des Unternehmens sollten diese ethischen Anforderungen dementsprechend gewichtet, operationalisiert und angepasst werden, um eine effektive und verantwortungsvolle Implementierung von KI-Systemen zu gewährleisten.

5. Zusammenfassung und Fazit

KI-Systeme bieten vielfältige Potenziale für Unternehmen und werden bereits heute in immer mehr Bereichen des unternehmerischen Alltags eingesetzt. Damit Unternehmen die vorhandenen Potenziale für sich nutzen können, müssen die unterschiedlichen Risiken, die mit KI-Systemen verbunden sind, minimiert werden. Genau hier setzen ethische Konzepte zu KI-Systemen an: Sie sollen helfen und Orientierung geben, KI-Systeme sicherer, transparenter und nutzerfreundlicher zu entwickeln und einzusetzen. Die Vielzahl ethischer Konzepte und Leitlinien, die in den letzten Jahren entstanden sind, kann jedoch für Unternehmen ein unübersichtliches und komplexes Feld darstellen.

Dieses Whitepaper verwendet das Konzept der *vertrauenswürdigen KI* der EK HLEG als Orientierungsrahmen und setzt es mit den Konzepten der *erklärbaren KI, ethischen KI, fairen KI, grünen KI, menschenzentrierten KI, und verantwortlichen KI* in Bezug. Die jeweiligen Anforderungen wurden mit den sieben Hauptanforderungen des Konzepts der vertrauenswürdigen KI abgeglichen. Dabei ergeben sich große Überschneidungen zwischen den analysierten Konzepten – die Unterschiede ergeben sich durch unterschiedliche Schwerpunktsetzungen innerhalb der Anforderungen.

Die Hauptanforderungen des Konzepts *vertrauenswürdiger KI* der EK HLEG (1) Vorrang menschlichen Handelns und menschliche Aufsicht, (2) Technische Robustheit und Sicherheit, (3) Datenschutz, Datenqualitätsmanagement und Privatsphäre (4) Transparenz, (5) Vielfalt, Nicht-Diskriminierung und Fairness, (6) gesellschaftliches und ökologisches Wohlbefinden sowie (7) Rechenschaftspflicht erweisen sich – im Abgleich mit den parallel entstandenen anderen Konzepten wie ethische oder verantwortliche KI – als gute und umfassende Grundlage, um sich mit ethischen Anforderungen an KI-Systeme zu beschäftigen.

Trotz der Einigkeit über die Relevanz der Erfüllung ethischer Leitlinien und vertrauenswürdiger KI für Unternehmen und der großen Überschneidungen zwischen den Konzepten, existieren Herausforderungen, die die Umsetzung für Unternehmen erschweren. Zusammengefasst sind diese Herausforderungen und konkrete Handlungsempfehlungen in Tabelle 3.

Vertrauenswürdige KI und ähnliche Konzepte wurden mit dem Anspruch entwickelt, für alle KI-Systeme und Anwendungskontexte gültig zu sein. Diese Allgemeingültigkeit bedeutet, dass vertrauenswürdige KI als Orientierung für Unternehmen dienen kann, um vertrauenswürdige KI-Systeme zu entwickeln und einzusetzen. Allerdings müssen Unternehmen diesen konzeptionellen Rahmen spezifisch für ihre eigenen Anwendungsfälle anpassen und operationalisieren. Das beinhaltet die Übersetzung und Quantifizierung der Anforderungen an vertrauenswürdige KI, das Abwägen dieser Anforderungen gegen die Leistungsfähigkeit der Systeme, das Festlegen klarer Verantwortlichkeiten sowie die Förderung von Partizipation und Mitbestimmung der Mitarbeitenden.

Tabelle 3: Überblick Herausforderungen und Handlungsempfehlungen für die Umsetzung der Anforderungen für vertrauenswürdige KI-Systeme.

Herausforderung	Handlungsempfehlung
Die Konzepte sind abstrakt und wenig handlungsleitend.	Vertrauenswürdige KI ist ein Grundgerüst. Bei einem spezifischen Anwendungsfall im Unternehmen kann es als Orientierung dienen. Dann müssen und können die Anforderungen an vertrauenswürdige KI im Unternehmen für den konkreten Fall spezifisch operationalisiert werden.
Der ethische und technische Diskurs verläuft getrennt und wird nicht zusammengedacht.	Interdisziplinär besetzte Teams in der Entwicklung und Implementierung von KI-Systemen in Unternehmen ermöglichen, dass ethische Anforderungen und technische Entwicklung zusammengedacht werden.
Anforderungen an vertrauenswürdige KI stehen im Spannungsverhältnis zur Leistungsfähigkeit sowie untereinander.	Im konkreten Anwendungsfall muss jeweils unternehmensindividuell zwischen den jeweiligen Anforderungen und in Bezug auf Leistungsfähigkeit abgewogen werden – dabei muss das spezifische Risiko für den jeweiligen Kontext bedacht werden.
Vermenschlichung von KI-Systemen führt zu Verantwortungsdiffusion.	Verantwortlichkeiten innerhalb des Unternehmens müssen vor und während der Entwicklung und des Einsatzes von KI-Systeme eindeutig geklärt sein.
Im Konzept der vertrauenswürdigen KI fehlt die Anforderung Partizipation.	Partizipation der Mitarbeitenden bei der Einführung von KI-Systemen sollte eingeplant werden – dabei muss Mitbestimmung und Raum für Bedenken sichergestellt werden.

KI-Systeme verantwortungsvoll zu entwickeln und einzusetzen und damit die identifizierten Herausforderungen zu überwinden, ist notwendig, um die Potenziale von KI-Systemen im Unternehmenskontext nutzbar zu machen. Unternehmen sollten sich daher aktiv mit dem Thema auseinandersetzen und Leitlinien konsequent für sich operationalisieren und umsetzen. Dies bietet nicht nur die Möglichkeit, Potenziale von KI-Systemen voll auszuschöpfen, sondern stärkt auch das nachhaltige Vertrauen in die Technologie. Mit dem richtigen Ansatz können Unternehmen eine Vorreiterrolle in der Nutzung vertrauenswürdiger KI übernehmen und gleichzeitig ethische sowie gesellschaftliche Standards voranbringen.

6. Literaturverzeichnis

Ameen, Nisreen; Tarhini, Ali; Reppel, Alexander; Anand, Amitabh (2021): Customer experiences in the age of artificial intelligence. In: Computers in Human Behavior 114, S. 106548. DOI: 10.1016/j.chb.2020.106548.

Angelov, Plamen P.; Soares, Eduardo A.; Jiang, Richard; Arnold, Nicholas I.; Atkinson, Peter M. (2021): Explainable artificial intelligence: an analytical review. In: Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 11 (5), Artikel e1424, e1424. DOI: 10.1002/widm.1424.

Barocas, Solon; Selbst, Andrew D. (2016): Big Data's Disparate Impact. In: SSRN Journal. DOI: 10.2139/ssrn.2477899.

Barredo Arrieta, Alejandro; Díaz-Rodríguez, Natalia; Del Ser, Javier; Bennetot, Adrien; Tabik, Siham; Barbado, Alberto et al. (2020): Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. In: Information Fusion 58, S. 82–115. DOI: 10.1016/j.inffus.2019.12.012.

Benbasat, Izak; Wang, Weiquan (2005): Trust In and Adoption of Online Recommendation Agents. In: Journal of the Association for Information Systems 6 (3), S. 72–101. DOI: 10.17705/1jais.00065.

Brundage, Miles; Avin, Shahar; Wang, Jasmine; Belfield, Haydn; Krueger, Gretchen; Hadfield, Gillian et al. (2020): Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims.

Capel, Tara; Brereton, Margot (2023): What is Human-Centered about Human-Centered AI? A Map of the Research Landscape. In: Albrecht Schmidt (Hg.): Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems. Unter Mitarbeit von Kaisa Väänänen, Tesh Goyal, Per Ola Kristensson, Anicia Peters, Stefanie Mueller, Julie R. Williamson und Max L. Wilson. CHI ,23: CHI Conference on Human Factors in Computing Systems. Hamburg Germany, 23 04 2023 28 04 2023. New York,NY,United States: Association for Computing Machinery (ACM Digital Library), S. 1–23.

Chamola, Vinay; Hassija, Vikas; Sulthana, A. Razia; Ghosh, Debshishu; Dhingra, Divyansh; Sikdar, Biplab (2023): A Review of Trustworthy and Explainable Artificial Intelligence (XAI). In: Institue of Electrical and Electronics Engineers Access 11, S. 78994–79015. DOI: 10.1109/ACCESS.2023.3294569.

Dastin, Jeffrey (2018): Insight - Amazon scraps secret AI recruiting tool that showed bias against women. In: Reuters Media, 11.10.2018. Online verfügbar unter https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G/, zuletzt geprüft am 12.02.2024.

Dwivedi, Rudresh; Dave, Devam; Naik, Het; Singhal, Smiti; Omer, Rana; Patel, Pankesh et al. (2023): Explainable AI (XAI): Core Ideas, Techniques, and Solutions. In: ACM Comput. Surv. 55 (9), S. 1–33. DOI: 10.1145/3561048.

EC HLEG on AI (2019): Ethik-Leitlinien für eine vertrauenswürdige KI. Europäische Kommission. Brüssel.

European Commission (2021): Regulation of the European Parliament and of the Council. Online verfügbar unter https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206, zuletzt aktualisiert am 14.02.2024, zuletzt geprüft am 14.02.2024.

European Parliament (2023): EU AI Act: first regulation on artificial intelligence. The use of artificial intelligence in the EU will be regulated by the AI Act, the world's first comprehensive AI law. Find out how it will protect you. Online verfügbar unter https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence, zuletzt aktualisiert am 19.12.2023, zuletzt geprüft am 12.02.2024.

Feuerriegel, Stefan; Dolata, Mateusz; Schwabe, Gerhard (2020): Fair AI. Challenges and Opportunities. In: Business & Information Systems Engineering 62 (4), S. 379–384. DOI: 10.1007/s12599-020-00650-3.

Freiman, Ori (2023): Making sense of the conceptual nonsense 'trustworthy AI'. In: AI and Ethics 3 (4), S. 1351–1360. DOI: 10.1007/s43681-022-00241-w.

Gillin, Paul (2021): Tackling Data Center Water Usage Challenges Amid Historic Droughts, Wildfires. In: Data Center Frontier, 08.01.2021. Online verfügbar unter https://www.datacenterfrontier.com/special-reports/article/11428474/tackling-data-center-water-usage-challenges-amid-historic-droughts-wildfires, zuletzt geprüft am 23.04.2024.

Hagendorff, Thilo (2020): The Ethics of AI Ethics: An Evaluation of Guidelines. In: Minds and Machines 30 (1), S. 99–120. DOI: 10.1007/s11023-020-09517-8.

Jobin, Anna; Ienca, Marcello; Vayena, Effy (2019): The global landscape of AI ethics guidelines. In: Nature Machine Intelligence 1 (9), S. 389–399. DOI: 10.1038/s42256-019-0088-2.

Kaur, Davinder; Uslu, Suleyman; Rittichier, Kaley J.; Durresi, Arjan (2022): Trustworthy Artificial Intelligence: A Review. In: Association for Computing Machinery - Computing Surveys 55 (2), S. 1–38. DOI: 10.1145/3491209.

Knight, Will (2017): The Dark Secret at the Heart of AI. In: Massachusetts Institute of Technology - Technology Review, 11.04.2017. Online verfügbar unter https://www.technologyreview.com/2017/04/11/5113/the-dark-secret-at-the-heart-of-ai/, zuletzt geprüft am 14.02.2024.

Lee, John D.; See, Katrina A. (2004): Trust in automation: designing for appropriate reliance. In: Human factors 46 (1), S. 50–80. DOI: 10.1518/hfes.46.1.50_30392.

Lukyanenko, Roman; Maass, Wolfgang; Storey, Veda C. (2022): Trust in artificial intelligence: From a Foundational Trust Framework to emerging research opportunities. In: Electron Markets 32 (4), S. 1993–2020. DOI: 10.1007/s12525-022-00605-4.

Mcknight, D. Harrison; Carter, Michelle; Thatcher, Jason Bennett; Clay, Paul F. (2011): Trust in a specific technology. In: Association for Computing Machinery - Transactions on Management Information Systems 2 (2), S. 1–25. DOI: 10.1145/1985347.1985353.

Mittelstadt, Brent (2019): Principles alone cannot guarantee ethical AI. In: Nature Machine Intelligence 1 (11), S. 501–507. DOI: 10.1038/s42256-019-0114-4.

Ozmen Garibay, Ozlem; Winslow, Brent; Andolina, Salvatore; Antona, Margherita; Bodenschatz, Anja; Coursaris, Constantinos et al. (2023): Six Human-Centered Artificial Intelligence Grand Challenges. In: International Journal of Human–Computer Interaction 39 (3), S. 391–437. DOI: 10.1080/10447318.2022.2153320.

Pieters, Wolter (2011): Explanation and trust: what to tell the user in security and AI? In: Ethics Inf Technol 13 (1), S. 53–64. DOI: 10.1007/s10676-010-9253-3.

Reinhardt, Karoline (2023): Trust and trustworthiness in AI ethics. In: AI Ethics 3 (3), S. 735–744. DOI: 10.1007/s43681-022-00200-5.

Schmitt, Lewin (2022): Mapping global AI governance: a nascent regime in a fragmented landscape. In: AI and Ethics 2 (2), S. 303–314. DOI: 10.1007/s43681-021-00083-y.

Shivdas, Sanjana; Kelly, Tim (2021): Toyota halts all self-driving e-Palette vehicles after Olympic village accident. Online verfügbar unter https://www.reuters.com/business/autos-transportation/toyota-halts-all-self-driving-e-pallete-vehicles-after-olympic-village-accident-2021-08-27/, zuletzt geprüft am 12.02.2024.

Siau, Keng; Wang, Weiyu (2020): Artificial Intelligence (AI) Ethics: Ethics of AI and Ethical AI. In: Journal of Database Management. Online verfügbar unter https://www.semanticscholar.org/paper/Artificial-Intelligence-%28AI%29-Ethics%3A-Ethics-of-AI-Siau-Wang/39d1f020a585d3f28cb4b4c14497649e6a469ef1.

Stahl, Bernd Carsten; Leach, Tonii (2023): Assessing the ethical and social concerns of artificial intelligence in neuroinformatics research: an empirical test of the European Union Assessment List for Trustworthy AI (ALTAI). In: AI and Ethics 3 (3), S. 745–767. DOI: 10.1007/s43681-022-00201-4.

Thiebes, Scott; Lins, Sebastian; Sunyaev, Ali (2020): Trustworthy artificial intelligence. In: Electron Markets 31 (2), S. 447–464. DOI: 10.1007/s12525-020-00441-4.

Verdecchia, Roberto; Sallou, June; Cruz, Luís (2023): A systematic review of Green AI. In: Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 13 (4), Artikel e1507. DOI: 10.1002/widm.1507.

Weber, Sebastian; Guldner, Achim; Begic Fazlic, Lejla; Dartmann, Guido; Naumann, Stefan (2023): Sustainability in Artificial Intelligence - Towards a Green AI Reference Model.

Weidinger, Laura; Uesato, Jonathan; Rauh, Maribeth; Griffin, Conor; Huang, Po-Sen; Mellor, John et al. (2022): Taxonomy of Risks posed by Language Models. In: 2022 ACM Conference on Fairness, Accountability, and Transparency. FAccT ,22: 2022 ACM Conference on Fairness, Accountability, and Transparency. Seoul Republic of Korea, 21 06 2022 24 06 2022. New York,NY,United States: Association for Computing Machinery (ACM Digital Library), S. 214–229.

Impressum

Fraunhofer-Institut für Arbeitswirtschaft und Organisation IAO
Nobelstraße 12
70569 Stuttgart
www.iao.fraunhofer.de

Kontakt
Jessica Wulf
Wissenschaftliche Mitarbeiterin
Gesellschaftliche Trends und Technologie
Mobil +49 151 67733047
jessica.wulf@iao.fraunhofer.de

Mitarbeit
Franziska Bolte
Ali Choukair

Satz und Layout
Valentin Buhl, Franz Schneider, Fraunhofer IAO

Titelbild
© elenabsl - Adobe Stock

Fraunhofer-Publica
http://dx.doi.org/10.24406/publica-3515

Kontakt

Jessica Wulf
Wissenschaftliche Mitarbeiterin
Gesellschaftliche Trends und Technologie
Mobil +49 151 67733047
jessica.wulf@iao.fraunhofer.de

Fraunhofer-Institut für Arbeitswirtschaft
und Organisation IAO
Nobelstraße 12
70569 Stuttgart

www.iao.fraunhofer.de