Zeitschrift für Praktische Philosophie
Band 11, Heft 1, 2024, S. 343–366
www.praktische-philosophie.org
https://doi.org/10.22613/zfpp/11.1.14

Die Verlagerung von Vertrauen vom Menschen zur Maschine: Eine Erweiterung des zwischenmenschlichen Vertrauensparadigmas im Kontext Künstlicher Intelligenz

The Shift of Trust from Human to Machine: An Expansion of the Interpersonal Trust Paradigm in the Context of Artificial Intelligence

Christopher Koska, München, Julian Prugger, München, Sophie Jörg, München & Michael Reder, München

Zusammenfassung: Der Beitrag untersucht die Transformation des Vertrauensbegriffs durch Künstliche Intelligenz. In aller Regel ist das Vertrauen in technische Artefakte eng mit deren Verlässlichkeit verbunden. Während diese Form der (technischen) Verlässlichkeit letztlich zumeist auf ein menschliches Gegenüber (z. B. auf die Herstellenden, die Betreibenden oder die Auditierenden) zurückgeführt wird, erfordert die zunehmende Präsenz von KI-Systemen eine Neubetrachtung dieses Zusammenhangs. Insbesondere bei selbstlernenden Systemen (wie der konnektivistischen KI), die ihre Auswahlkriterien und Selektionsmechanismen durch die Interaktion mit der Umwelt kontinuierlich verändern, zeichnet sich eine graduelle Verschiebung des Vertrauens ab: Vertrauen wandert, so lautet die These im Anschluss an Hartmann (2022), Schritt um Schritt vom Menschen zur Maschine. Der Beitrag zielt auf eine problemorientierte Beschreibung der damit verbundenen Chancen und Herausforderungen.[1]

1 Dieser Artikel ist Ergebnis mehrerer Forschungsprojekte, insbes. des vom bidt geförderten Projekts „Kann ein Algorithmus moralisch kalkulieren? (KAIMo)“ und des Projekts „medAIcine“, dem Pilotprojekt des von der Hochschule für Philosophie, der TUM und der Uni Augsburg neu gegründeten „Centers for Responsible AI Technologies (CReAITech)“ und des von der DFG geförderten Projektes Politics in Search of Evidence (PoSEvi, Projektnummer AP 235/6-

Schlagwörter: Mensch-Maschine-Interaktion, Mehr als Verlässlichkeit, Eigendynamik von KI

Abstract: The article investigates the transformation of the concept of trust due to Artificial Intelligence. Traditionally, trust in technical artifacts is closely linked to their reliability. While this form of (technical) reliability is usually traced back to a human counterpart (e.g., the manufacturers, operators, or auditors), the increasing presence of AI systems necessitates a reconceptualisation of this relationship. Particularly in self-learning systems (like connectivist AI), which continuously modify their selection criteria and mechanisms through interaction with the environment, a gradual shift in trust is becoming apparent: The thesis posits that trust shifts, following Hartmann (2022), incrementally from humans to machines. The article aims to provide a problem-oriented description of the associated opportunities and challenges.

Keywords: Human-Machine Interaction, More than Reliability, Normative Momentum of AI

1 Einleitung

Schon seit einigen Jahren überschlagen sich Nachrichten über technologische Fortschritte durch den Einsatz von KI-basierten Werkzeugen in ganz unterschiedlichen Feldern: insbesondere im Kontext des autonomen Fahrens oder mit Blick auf den Einsatz von KI in der Medizin bis hin zur Weiterentwicklung von Sprachmodellen für Chatbots (wie Google LaMDA oder ChatGPT). Im Zuge dieser Entwicklungen wurde in den vergangenen Jahren vermehrt über das Vertrauen in die neuen Technologien diskutiert.

Vertrauen wird in den Diskussionen über KI unterschiedlich thematisiert: Vertrauen wird z. B. problematisiert, wenn politische Institutionen oder Unternehmen Daten in großen Mengen sammeln und diese für die Entwicklung künstlicher Intelligenz nutzen. Angefangen von der globalen NSA-Affäre im Sommer 2013 (Greenwald 2014) über die Auswertung von Big Data im politischen Wahlkampf (Grassegger und Krogerus 2016) bis hin zu den Ambitionen einiger Tech-Giganten, menschliche Gedanken unmittelbar auszulesen (Rixecker 2017; Meckel 2018; Nordenbrock 2022), wird dabei zumeist ein Machtungleichgewicht, das Eindringen in die Privatsphäre oder eine übermäßige Kontrolle (durch den Staat oder das Unternehmen) kritisiert. Zudem wird das Vertrauen in Frage gestellt, wenn Maschinen in bestimmten

1), das mit der Universitätsmedizin der Universität Magdeburg (Team Christian Apfelbacher) durchgeführt wird.

Bereichen deutlich leistungsfähiger werden als der Mensch und spezifische menschliche Kompetenzen durch Künstliche Intelligenz ersetzt werden.

Die Modellierung von KI baut auf bestimmten Aspekten der menschlichen Kognition auf und lässt sich darüber in zwei Hauptformen unterteilen (Russell und Norvig 2022): Die symbolverarbeitende oder auch klassische KI modelliert menschliches Denken durch die Kodierung von Wissen und logischen Regeln. Diese Form von KI eignet sich vor allem für den Einsatz in Feldern, in denen ein Verständnis von logischen Formen nötig und klar definierte Regeln wichtig sind. Auf der anderen Seite modelliert die konnektivistische KI menschliches Lernen und Mustererkennung durch die Simulation von neuronalen Netzwerken. In der Praxis wird aufgrund der besseren Performance und Flexibilität häufig eine Kombination von symbolverarbeitenden und konnektivistischen Ansätzen verwendet.

In der medizinischen Diagnostik werden beispielsweise konnektivistische Formen von KI genutzt, um Muster in medizinischen Bildern oder Patient*innendaten zu erkennen. Gleichzeitig werden symbolverarbeitende Ansätze verwendet, um medizinisches Wissen zu kodieren, zu nutzen und Diagnosen zu unterstützen. Auch bei selbstfahrenden Fahrzeugen wird aus Gründen der Effizienz häufig eine Kombination aus symbolverarbeitenden und konnektivistischen Ansätzen eingesetzt. Hier können symbolverarbeitende Systeme z. B. verwendet werden, um präzise Karten zu erstellen und Routenplanung durchzuführen, während konnektivistische Ansätze (wie Deep Learning) für Aufgaben der Objekt- und Verkehrszeichenerkennung verwendet werden. Insbesondere durch die Entwicklungen im Feld konnektivistischer KI und der Integration von klassischen Formen wird die Diskussionen über Vertrauen und KI gegenwärtig sehr kontrovers geführt (Reinhardt 2022; Friedrich, Seifert und Schleidgen 2023). Aus philosophischer Sicht ist es angesichts dieser teils stark zugespitzten Kontroverse wichtig zu klären, was gemeint ist, wenn von Künstlicher Intelligenz gesprochen wird.

Dieser Beitrag versteht KI-Systeme als integralen Bestandteil von medialen Werkzeugen (Manovich 2005, 2018). Während traditionelle technische Werkzeuge von den Nutzenden meist aktiv ausgewählt und genutzt werden, um spezifische Bedürfnisse und Ziele zu erfüllen (bspw. Empfehlungen oder konkrete Informationen, soziale Interaktionen), findet bei den KI-Technologien in aller Regel keine Auswahl durch die Mediennutzenden, sondern durch die Medienschaffenden, statt. Wenn von Vertrauen in KI-Systeme gesprochen wird, unterscheiden wir deshalb nicht zwischen symbolverarbeitender und konnektivistischer KI. Bei einem nutzerzentrierten Zugang

steht die Technik – verstanden als ein mediales Werkzeug – als Ganzes im Zentrum der Betrachtung, zumal in der Praxis oft beide Ansätze gleichzeitig zum Einsatz kommen. Die Erklärbarkeit von KI-Systemen, insbesondere bei konnektivistischen Ansätzen und oft im Kontext von „xAI" (Explainable Artificial Intelligence) diskutiert, stellt jedoch eine besondere Herausforderung dar. Deshalb wird im weiteren Text auch erörtert, inwiefern Maßnahmen wie ein Algorithmen-TÜV und die Mediatisierung von zwischenmenschlichen Vertrauensbeziehungen einen positiven Einfluss auf das Vertrauen der Mediennutzenden nehmen können.

Eine Frage des gegenwärtigen Diskurses hinsichtlich des Vertrauens ist, ob Vertrauen in KI mehr als Verlässlichkeit erfordert, die bei der Diskussion um Vertrauen in herkömmliche Technologien meist im Zentrum stand. Die zentrale Frage ist, wie sich dieses Mehr gegenüber der Verlässlichkeit theoretisch fassen lässt und welche Konsequenzen sich daraus für das Verhältnis von Menschen und KI-Systemen ergeben.

Zur Diskussion dieser Frage ist es wichtig, drei Implikationen des Diskurses zu differenzieren: *Erstens* wird aus philosophischer und soziologischer Perspektive Vertrauen primär als ein Sich-Verlassen auf ein Gegenüber definiert, insbesondere wenn das Ergebnis einer Handlung unsicher oder riskant ist und hierfür Kontrolle aufgeben werden muss (Gloyna 2017). Vertrauen kann beispielsweise an die rationale Erklärbarkeit gekoppelt werden. Als Voraussetzung für Verlässlichkeit wird dann häufig eine rationale, epistemisch begründbare Erwartungshaltung angeführt. Oder aber Vertrauen wird stärker mit einer emotionalen Komponente verschränkt (Jones 1996). Zudem kann Vertrauen auch als ein Wert interpretiert werden, der jenseits einer rationalen und emotionalen Verhältnissetzung auf der individuellen Ebene eine spezifische normative Forderung darstellt.

Zweitens wird Vertrauen philosophisch oft auf zwischenmenschliche Interaktionen zurückgeführt (McLeod 1999; Faulkner 2015; Hartmann 2020). Erst die Begegnung von Subjekten ermöglicht es, Vertrauen zu bilden, so die These. Ein solches intersubjektives Vertrauensparadigma kann technische Objekte vor allem als passive Empfänger oder Vermittler von Vertrauen erfassen. Vertrauen wird aus dieser Perspektive wiederum teilweise auf die Zuverlässigkeit technischer Systeme reduziert; auch um sich nicht der Kritik einer unangemessenen Vermenschlichung von KI-Technologien auszusetzen. Durch ein enggeführtes intersubjektives Vertrauensparadigma besteht jedoch die Gefahr, dass die Fähigkeit der KI übersehen wird, als aktives Element in der Vertrauensbildung zu fungieren.

Drittens haben feministische Autor*innen wie Baier (1986) argumentiert, Vertrauen bedeute zuzulassen, selbst potenziell geschädigt zu werden. Zu vertrauen sei deshalb nicht primär eine Frage des Willens oder der rationalen Zustimmung (und damit der Verlässlichkeit), sondern setzt ein soziales oder politisches Klima voraus, das es ermöglicht, sich verletzbar zu machen.

In alle drei Richtungen zeigt sich, so die These des Beitrags, dass Vertrauen in KI mehr ist als Verlässlichkeit und rationale Erwartbarkeit. Der Beitrag entfaltet entlang der drei Diskussionsstränge mögliche Linien eines solchen erweiterten Verständnisses von Vertrauen gegenüber KI-Systemen. Nach diesem einführenden ersten Teil wird im zweiten Teil ein Modell zur Differenzierung zwischen Vertrauensgebenden, Vertrauensobjekt und Vertrauensnehmenden eingeführt, um die Verschränkung von rationalen, emotionalen und normativen Komponenten von Vertrauen mit Blick auf die Verlässlichkeit von KI-Systemen zu skizzieren. Im dritten Abschnitt wird das zwischenmenschliche Vertrauensparadigma kritisch hinterfragt und die Verlagerung von Vertrauen vom Menschen zur Maschine thematisiert. Zudem wird die Verbindung von KI und Selbst problematisiert und das Phänomen der inhärenten Normativität der Digitalität skizziert. Im vierten Abschnitt wird daran anknüpfend die Dimension der Verletzbarkeit aus einer poststrukturalistischen Perspektive als Voraussetzung für Vertrauen reflektiert. Abschließend werden im fünften Teil die Ergebnisse noch einmal zusammengefasst und zwei Ausblicke auf zukünftige Forschungsdesiderate formuliert.

2 Verlässlichkeit und rationale Erwartungshaltung als Ausgangspunkt

Verlässlichkeit und rationale Erwartungshaltungen stellen zwei wichtige Faktoren für Vertrauen dar. Diese Erkenntnis gründet in der Überlegung, dass Vertrauen weitgehend auf dem Glauben an die Erfüllung bestimmter Vorhersagen oder Erwartungen beruht, welche sich aus vergangenen Erfahrungen und vorliegenden Informationen speisen. Im Kontext von KI-Technologien können diese Erwartungen als rationale Erwartungshaltungen der Nutzenden, Betreibenden oder Herstellenden verstanden werden, während sich die Verlässlichkeit auf die Konsistenz der Leistung und die Ergebnisse von KI-Technologien bezieht. Im folgenden Abschnitt sollen zunächst die Elemente von Vertrauensbeziehungen differenziert werden, die auf der Ver-

lässlichkeit von KI-Systemen und rationalen Erwartungshaltungen beruhen. Dabei wird gezeigt, dass emotionale Aspekte bei Mediennutzenden eng mit rationalen Erwartungshaltungen verbunden sind. Zudem wird argumentiert, dass Zertifizierungsverfahren dazu beitragen können, spezifische normative Anforderungen an ‚vertrauenswürdige KI' (*trustworthy AI*) zu stellen, um das Verständnis von Vertrauen in KI – über den Aspekt der Verlässlichkeit hinaus – zu erweitern.

Vertrauensbeziehungen lassen sich als dreiteilige Relationen darstellen, wie Michalski (2019, 44) zeigt: „Ein Vertrauensgeber A vertraut einem Vertrauensnehmer B hinsichtlich eines Vertrauensobjektes X". Im Falle von KI sind die Nutzenden die Vertrauensgebenden, da sie den Herstellenden bzw. Betreibenden (den Vertrauensnehmenden) ihr Vertrauen mit Bezug auf ein konkretes KI-System (Vertrauensobjekt) schenken.

Unsicherheit wird minimiert, wenn die Funktionsweise einer bestimmten KI-Technologie den rationalen Erwartungen der Vertrauensgebenden entspricht. Verlässlichkeit ist insofern ein fundamentaler Baustein für das Vertrauen, weil sie den Nutzenden die Sicherheit gibt, dass das technische System so funktionieren wird, wie es soll. Dies ist wiederum nicht nur eine rationale Erkenntnis, sondern die Verlässlichkeit von KI-Systemen erzeugt auch auf der emotionalen Ebene ein Gefühl der Sicherheit. Menschen fühlen sich sicher, wenn sie sich auf technische Medien verlassen können (Becker 2018). Dies ist ein erster Hinweis darauf, dass Vertrauen nicht nur als im Sinne einer rationalen Erwartbarkeit (Verlässlichkeit) verstanden werden sollte, sondern dann in das Vertrauen auch emotionale Aspekte einfließen. Dies zeigt sich auch dann, wenn die Funktionsweise von technischen Medien nicht (mehr) den rationalen Erwartungen der Nutzenden (oder auch der Betreibenden usw.) entspricht.

Mit Blick auf KI-basierte Assistenzsysteme lässt sich in vielen Handlungskontexten zunächst feststellen, dass die Technologie nicht für alle Nutzenden gleichermaßen transparent sein kann. Denn Transparenz ist ein Relationsbegriff, der u. a. von dem Kenntnisstand der Nutzenden und dem jeweiligen Handlungskontext abhängt (Koska 2023, 153–155). Aus der Perspektive von Mediennutzenden, die KI-basierte Assistenzsysteme über adaptive Mensch-Maschine-Schnittstellen bedienen, ist es entscheidend, dass die für ihren spezifischen Handlungskontext relevanten Aspekte des Systems verständlich und nachvollziehbar sind (Martini 2019). Überzogene Transparenzanforderungen könnten den Mehrwert solcher Systeme dagegen schnell ad absurdum führen. Wenn die Mediennutzenden die Verlässlich-

keit der zugrundeliegenden KI-Technologien für jedes Anwendungsszenario im Detail verstehen müssten, könnte dies zu einer Informationsüberflutung führen und die Nutzungsfreundlichkeit beeinträchtigen (Ananny und Crawford 2016). Für Fragen der Verlässlichkeit und rationalen Erwartbarkeit geht es deshalb um die Frage, ob sich die Funktionsweise des KI-Systems kontextspezifisch für unterschiedliche Zielgruppen vermitteln lässt und die Grenzen der KI für unterschiedliche Anwendungsszenarien erkannt werden können. Denn dann lassen sich personale Autonomie, Datensouveränität und digitale Selbstbestimmung auf dem Abstraktionslevel verorten, auf dem Transparenz hergestellt werden kann.

Verlässlichkeit und rationale Erwartbarkeit als ein Baustein von Vertrauen müssen also differenziert gefasst werden, je nach Kontext und Ausgangslage der involvierten Personen. Dabei spielen nicht nur die Herstellenden und die Betreibenden eine zentrale Rolle. Weil die Ausgangslagen so unterschiedlich sind, wird in sozialen Praktiken (bspw. in ökonomischen Zusammenhängen) auf Institutionalisierungen zurückgegriffen, um damit die Verlässlichkeit der Systeme durch die Prüfung Dritter zum Ausdruck zu bringen. Die rational begründbare Erwartungshaltung der Nutzenden wird dann gezielt mit einer emotionalen Komponente verschränkt, die bestimmten Marken oder Labels (bspw. einem TÜV-Siegel) entgegengebracht wird. Eine rein technische Transparenz der Systeme reicht nicht aus, so die These, um ein tiefes Verständnis und Vertrauen bei den Mediennutzenden zu gewährleisten. Transparenzanforderungen müssen vielschichtig betrachtet werden: Während Expertenaudits ein tiefgehendes Verständnis der Systemfunktionsweise erfordern und wir in diesem Zusammenhang nicht sinnvoll von einer überzogenen Transparenzanforderung sprechen können, genügt für die allgemeinen Nutzenden eine verständliche und transparente Darstellung auf einer höheren Abstraktionsebene. Insbesondere in komplexen, konfliktiven und risikobehafteten Anwendungsfeldern steigt, wie u. a. der Entwurf des AI-Acts (European Commission 2021) zeigt, zugleich der Bedarf, Kriterien wie ‚Verlässlichkeit' und ‚Robustheit' durch weitere Dimensionen zu ergänzen. In diesem Beitrag plädieren wir daher im vierten Abschnitt für ‚Verletzbarkeit' als zusätzliche Dimension.

Für die Zertifizierung vertrauenswürdiger KI-Systeme gibt es bisher noch keine anerkannten, weltweit gültigen Labels. Kriterien-basierte Ansätze, wie das VCIO-Modell (VDE 2022) und prozessorientierte Modelle, wie der ISO/IEC/IEEE 24748-7000 (IEEE 2021), zeigen aber bereits auf, in welche Richtungen dabei gedacht wird (Wolf et al., im Druck). Das VCIO-Mo-

dell etwa zielt darauf ab, eine Vergleichbarkeit zwischen ähnlichen Produkten und Dienstleistungen herzustellen. Hierfür gibt es einen ausführlichen Kriterienkatalog mit fünf Werten (*Transparency, Accountability, Privacy, Fairness* und *Reliability*) speziell für KI-Systeme. Der IEEE-7000-Standard ist hingegen darauf ausgerichtet, ethische Werte (wie Vertrauen, Transparenz, Fairness u. v. m.) in das Design und die Implementierung von technischen Systemen, einschließlich KI-Systemen, zu integrieren. Dies geschieht durch die kontextspezifische Definition von sogenannten *Ethical Value Requirements (EVR)*, die dann in spezifische Produkt- und Organisationsanforderungen übersetzt werden. Beide Ansätze gehen davon aus, dass Vertrauen in KI mehr als Verlässlichkeit erfordert und dass wir für die rationale Analyse von Vertrauensobjekten konkrete Indikatoren (spezifische Kriterien oder Produkteigenschaften) benötigten. Speziell der prozessorientierte Ansatz des IEEE-7000-Standards ermöglicht dabei eine anwendungsbezogene Erweiterung von Vertrauen in KI, indem spezifische normative Forderungen (über die EVRs) bereits in das Design von KI-Systemen einfließen.

Die vorangegangenen Überlegungen haben gezeigt, dass Vertrauen mehr als eine rationale Erwartungshaltung ist, insofern Vertrauen auch eine emotionale und eine wertebasierte Komponente aufweist, was beides an Beispielen von Praktiken des Vertrauens aufgezeigt wurde. Vertrauen ist, so die Schlussfolgerung, immer schon mehr als Verlässlichkeit und impliziert dabei eine emotionale Form der Beziehung, auch wenn diese häufig vernachlässigt wird. Zertifizierung und Markenbildung sind mögliche Praktiken einer Vertrauensbildung, in denen beide Dimensionen von Vertrauen gleichermaßen angesprochen und umgesetzt werden.

3 Transformation des intersubjektiven Vertrauensparadigmas

In dem Modell der dreiteiligen Vertrauensbeziehung werden KI-Systeme als technische Objekte oder Vertrauensgegenstände erfasst. Diese Sichtweise spiegelt wider, dass Vertrauen zumeist als eine zwischenmenschliche Dynamik betrachtet wird, die auf persönlicher Erfahrung sowie sozialen Verständigungs- und Kommunikationsprozessen beruht. Aus einer solchen Perspektive wird Vertrauen als das Ergebnis einer Beziehung zwischen zwei oder mehreren Subjekten angesehen, bei der Vertrauen aufgebaut, aufrechterhalten oder verletzt werden kann. Die Annahme, dass technische Systeme, wie KI-Modelle, nicht als Vertrauenssubjekte betrachtet werden können, sondern nur als Vertrauensobjekte, findet sich bereits in den oben skizzierten

Zertifizierungsansätzen. Zuverlässigkeit, Sicherheit, Fairness, Transparenz usw. werden hier lediglich als Indikatoren für Vertrauen verstanden.

Allerdings deuten bereits die Zertifizierungsansätze daraufhin, dass das Vertrauen eher auf organisatorischen Prozessen und auf dem Vertrauen in die Herstellenden, Betreibenden und Auditierenden beruht. Zunächst gilt es zu betonen, dass das primäre Ziel von KI-Systemen nicht in erster Linie darin besteht, zwischenmenschliche Vertrauensbeziehungen zu etablieren. Stattdessen dienen sie vor allem dazu, spezifische Funktionen und Aufgaben in einer Vielzahl von Anwendungsbereichen zu erfüllen. Dennoch können KI-Systeme auch als Werkzeuge dienen, die von Menschen genutzt werden, um zwischenmenschliche Vertrauensbeziehungen zu mediatisieren, wobei sie dann zusätzlich als technisiertes Bindeglied zwischen Menschen fungieren. Ein solches Verständnis knüpft an die Gebrauchstheorie von Medien in der Tradition des späten Wittgenstein (Margreiter 2003, 151ff) an, die auch in landläufigen Aussagen wie „Die Technik als solches ist neutral!“, „Medien sind Werkzeuge“ und „Es kommt letztlich nur darauf an, wie wir Medien bzw. Werkzeuge gebrauchen“ ihren Ausdruck findet. Ausgehend von dem Verständnis, dass KI-Systeme Medien sind und Medien „Werkzeuge, die der Koordination zwischenmenschlichen Handelns dienen“ (Sandbothe 2003, 195), wird im Folgenden in einem ersten Schritt die Neutralitätsthese von Technik – speziell mit Blick auf KI-Systeme – kritisch hinterfragt. Aufbauend auf dieser Analyse wird in einem zweiten Schritt die graduelle Verschiebung von Vertrauen als eine sukzessive Verlagerung vom Subjekt zum Objekt des Vertrauens problematisiert.

KI-Systeme sind normativ, aber sie setzen sich selbst keine Normen. Sie sind normativ, insofern sie Regeln, Vorgaben und Standards befolgen, die bei ihrer Programmierung festgelegt wurden („Code is law“, wie Lessig bereits 1999 proklamierte). Diese Regeln können ethische Normen berücksichtigen (ISO/IEC/IEEE 24748-7000), gesetzliche Bestimmungen (EU-DSGVO) oder einfach nur programmtechnische Anforderungen sein. Die technischen Voraussetzungen wirken sich in einem weiteren Sinne normativ aus, da KI-Systeme – auf der Grundlage von Daten und Algorithmen – nur in einer digitalisierten Welt operieren können. Daten sind eine Abstraktion der Wirklichkeit, ein digitaler Schatten der Realität: Was digital nicht erfasst wird, existiert für KI-Systeme nicht (Koska 2023, 11–73).

Diese Ebene lässt sich auch als die ontologische Dimension der Digitalität bezeichnen (ebd., 45–48). KI-Systeme können sich nicht selbst in ein normativ-wollendes Verhältnis zu bestimmten Weltzuständen setzen. Das

liegt daran, dass sie über kein Bewusstsein und keine Selbstwahrnehmung verfügen, um eigene Wünsche, Ziele und Interessen zu verfolgen. Zwar gibt es KI-Technologien, sowohl aus dem Bereich der symbolischen oder klassischen KI (z. B. evolutionäre Algorithmen) als auch aus dem Bereich der konnektivistischen KI (z. B. neuronale Netze), welche die vom Menschen festgelegten Regeln und Parameter verändern können. Allerdings setzen sie sich selbst keine eigenen Ziele. Obwohl solche Systeme lernen und sich anpassen können, sind sie nicht in der Lage, ihre eigene Situation zu überdenken. Selbstlernende Systeme können sich nur innerhalb der Grenzen anpassen, die durch ihre Architektur, Trainingsdaten, Lernalgorithmen und Zielfunktion festgelegt wurden. Die Ziele von Optimierungsalgorithmen werden (bspw. beim Reinforcement Learning) von der Struktur des zu lösenden Problems und den vom Menschen festgelegten Belohnungen oder Strafen bestimmt. Vor diesem Hintergrund lassen sich KI-Systeme nach wie vor als Werkzeuge betrachten, die Menschen z. B. nutzen, um bestimmte Optimierungsprobleme zu lösen.

Die zunehmend komplexer werdende Interaktion zwischen Menschen und Maschinen erweitert jedoch die rein funktionale Dimension von KI-Systemen und konfrontiert uns mit grundlegenden Überlegungen zur Natur und dem Wesen von Vertrauen. Insbesondere der Wandel in der Wahrnehmung von Mediennutzenden führt zu einer Neubewertung der Beziehungen zwischen Menschen und Maschinen im Kontext von Vertrauen. In diesem Beitrag argumentieren wir dafür, technische Objekte nicht nur als passive Empfänger oder Vermittler von Vertrauen zu erfassen, sondern auch über die Fähigkeit von KI-Systemen nachzudenken, als aktives Element in der Vertrauensbildung zu fungieren. Im Kontext von KI-Systemen, die immer komplexere Aufgaben ausführen und Entscheidungen treffen, die Menschen nicht direkt kontrollieren können, verlagert sich Vertrauen langsam von den Menschen zu den Maschinen. Dabei geht es, wie Martin Hartmann anhand eines Beispiels aus der Medizintechnik skizziert, gar nicht darum, Vertrauen abzuschaffen, sondern um die Verlagerung von Vertrauen aus der Perspektive der Mediennutzenden: Hartmann berichtet von einem Vorfall in einem Zug, bei dem ein Arzt einer kollabierten Frau mit Wiederbelebungsmaßnahmen zu Hilfe kam. Trotz der dringlichen Situation wies eine mitreisende ältere Dame den Arzt wiederholt darauf hin, er solle genau das tun, was der sprechende Defibrillator vorschlägt, und offenbarte so ein stärkeres Vertrauen in die Maschine als in den Menschen. Diese Situation veranschaulicht, wie in einem kritischen Moment das Vertrauen nicht nur beim Mediziner,

sondern auch bei der Technologie lag. „Was ich daran jetzt interessant fand, ist, dass das Vertrauen sich hier verschoben hat. Wir können jetzt lange darüber reden, ob es Maschinenvertrauen gibt […] Für ihn [Anmerkung: den Arzt] ist aber besonders relevant, dass sozusagen das Vertrauen gewandert ist – hin zur Technik" (Hartmann 2022, 46:05). Hartmann folgert aus diesem Beispiel, dass die Verschiebung des Vertrauens nicht notwendigerweise zu einem Verlust an Vertrauen führt, sondern eher einer Neuzuweisung von Vertrauen in Richtung Technologie entspricht. Unabhängig von der philosophischen Frage, ob es Maschinenvertrauen gibt oder nicht, verdeutlicht diese Implikation bereits eine Anpassung an die neuen, technologisch geprägten Vertrauensdynamiken unserer Gesellschaft.

Durch Maschinen, die eigenständig lernen und sich (im Rahmen der oben abgesteckten Grenzen) weiterentwickeln, wird die Frage aufgeworfen, inwiefern traditionelle Indikatoren für Vertrauen, wie Verlässlichkeit und rationale Erwartungshaltung, durch andere Aspekte ersetzt oder ergänzt werden. Angesichts der steigenden Komplexität von KI-Systemen und der Forderung nach einer für Mediennutzende relevanten Transparenz, ist oft weniger das tiefgehende technische Verständnis der KI entscheidend, als vielmehr das Verstehen ihrer wesentlichen Funktionen und Auswirkungen. In Situationen, in denen die spezifische Funktionsweise der KI durch die „Kulturtechnik der Verflachung" (Krämer 2020) hinter einem kontextspezifischen Abstraktionsniveau verborgen bleibt, läuft man allerdings Gefahr, dass die Leistungsfähigkeit – also die Fähigkeit des Systems, spezifische Herausforderungen effektiv zu bewältigen – zum zentralen Kriterium in den Abwägungs- und Entscheidungsprozessen wird. Diese einseitige Fokussierung auf die Leistungsfähigkeit kann dazu führen, dass Nutzende Risken unterschätzen und ihre Entscheidungen auf simplifizierten Sichtweisen begründen, ohne sich eingehend mit den möglichen Grenzen und vielschichtigen Konsequenzen des KI-Einsatzes auseinanderzusetzen.

Als Mediennutzende befinden wir uns bereits in vielen Handlungskontexten an einem Kipppunkt, wenn die Funktionsweise des KI-Systems das Verständnis menschlicher Akteure übersteigt. In solchen Situationen kann das Vertrauen in die Leistungsfähigkeit (Fähigkeit des Systems, bestimmte Aufgaben und Probleme zu lösen) zum maßgeblichen Kriterium in den Abwägungs- und Entscheidungsprozessen von Mediennutzenden werden. Doch was bedeuten Verlässlichkeit und Leistungsfähigkeit genau und inwiefern ist es ratsam, die beiden Begriffe zu unterscheiden, aber nicht voneinander zu trennen? Verlässlichkeit bezeichnet die konsistente und vor-

hersehbare Performance eines Systems, während Leistungsfähigkeit dessen Fähigkeit ist, Aufgaben qualitativ hochwertig und effizient zu erfüllen. Ein leistungsstarkes System kann dennoch unzuverlässig sein. Bis zu einem gewissen Schwellenwert ist die Verlässlichkeit aber immer an die Leistungsfähigkeit gekoppelt. Ein KI-System, das bspw. 95% der Zeit beeindruckende Genauigkeit zeigt, aber in 5% der Anwendungen gravierende Abweichungen aufweist, kann in sicherheitskritischen Anwendungen wie der Medizintechnik verheerende Folgen haben. In solchen Kontexten wird die Bedeutung von Verlässlichkeit umso deutlicher und sollte nicht von der Leistungsfähigkeit überschattet werden.

Ein weiterer kritischer Faktor mit Blick auf die Verlässlichkeit ist die potenzielle Diskrepanz zwischen der tatsächlichen Arbeitsweise der Maschine und den ursprünglichen Absichten der Entwickelnden. Als Ursache wird häufig angeführt, dass die Nutzenden das Medium (Werkzeug) falsch oder anders benutzt haben, als es von den Herstellenden intendiert wurde. Ein bekanntes Beispiel hierfür, sind die rassistischen Entgleisungen des Microsoft Chatbots *Tay*: „Nutzer brachten Tay unter anderem dazu, Adolf Hitler zu preisen, den Holocaust zu verneinen und Schwarze zu beleidigen“ (Heise 2016). Solche Beispiele lassen sich aber nur begrenzt unter Verweis auf die *Product Misuse* oder *Dual Use* Problematik erklären, die selbstverständlich für alle Arten von Werkzeugen gilt. Friedmann und Nissenbaum haben schon 1996 auf Verzerrungseffekte in Computersystemen aufmerksam gemacht, die über präexistierende gesellschaftliche Strukturen, Einstellungen und Praktiken hinausweisen.[2] Diese Beobachtungen verweisen auf ein Phänomen, welches wir als inhärente Normativität der Digitalität bezeichnen. Aufgrund der fortschreitenden Automatisierung lässt sich in diesem Zusammenhang eine technologiebedingte Eigendynamik beobachten, die auf die kollektive Dimension von Algorithmen verweist (Jaume-Palasí und

2 Friedman und Nissenbaum (1996, 333–336) unterscheiden drei Hauptkategorien für Verzerrungseffekte (Bias) in Computersystemen: i) „Preexisting Bias“: bereits bestehende Verzerrungen, die ihre Wurzeln in präexistenten Strukturen, Einstellungen und Praktiken haben und sich dann auch in den algorithmischen System wiederfinden lassen, ii) „Technical Bias“: Verzerrungseffekte, die auf technischen Erwägungen, Einschränkungen oder Zwängen beruhen, iii) „Emergent Bias“: entstehende bzw. neuaufkommende Verzerrungseffekte, die im Zusammenhang mit der praxisbezogenen Systemanwendung von Personen entstehen (dieser Verzerrungseffekt entsteht typischerweise einige Zeit nach der Fertigstellung eines Systems als Folge sich ändernder gesellschaftlicher Kenntnisse, Fähigkeiten oder kultureller Werte).

Spielkamp 2017, Koska 2023, 112–119). Das Phänomen der sozialen Konstruktion ist hierfür ein gutes Beispiel. Es beschreibt, wie KI-Systeme u. a. im Online-Marketing eingesetzt werden, um Identitätsmerkmale (Ziele, Interessen und Bedürfnisse) von außen zu organisieren. Insofern dabei keine Rücksicht auf die Person genommen wird, der bestimmte Merkmale durch Dritte zugeschrieben werden, geht es bei der Personalisierung und Kontextualisierung aber weniger um die Befriedigung individueller Interessen und Bedürfnisse, als vielmehr darum, individuelle Wünsche und Ziele in der „numerischen Allgemeinheit" (Heesen 2018) von Affinitätsmodellen und Anonymisierungsmethoden aufzulösen. In der Praxis verfügen die Mediennutzenden (trotz Einführung der EU-DSGVO) nur über äußerst begrenzte Verfügungs- und Steuerungsmöglichkeiten bezüglich ihrer Datenprofile, so dass sich der digitale Zwilling, wie Yvonne Hofstetter es formuliert, zumeist als „virtueller Zombie" offenbart.

Diese inhärente Normativität von KI-Systemen, die Identitätsmerkmale organisiert, bleibt für KI-nutzende Personen dabei nicht immer rein äußerlich. Das bedeutet, KI-Nutzende nehmen die Identitätsmerkmale, die von KI an sie herangetragen werden, nicht immer als von außen konstruierte Identitäten wahr, von welchen sie sich selbst leicht abgrenzen könnten. Im folgenden Abschnitt soll mit Hilfe eines Perspektivwechsels dargestellt werden, wie im Anschluss an ein poststrukturalistisch-feministisches Subjektverständnis Verletzbarkeit als quasi anthropologische Grundkonstante dazu führt, dass teilweise nicht scharf zwischen einer inhärenten Normativität der KI und dem Selbst der Nutzenden unterschieden werden kann. Dies hat wiederum Auswirkungen auf die Bedeutung von Vertrauen für KI-Kontexte.

4 Verbindung von KI und Selbst: Verletzbarkeit als Voraussetzung für Vertrauen

Die bisherigen Überlegungen haben gezeigt, dass das Vertrauen, mit Hartmann gesprochen, vom Menschen zur Maschine wandert und wie sich damit auch traditionelle Kriterien wie Verlässlichkeit und rationale Erwartbarkeit verändern. Die Auswirkung dieser ‚Verwobenheit' von Mensch und KI-Systemen soll im letzten Schritt aus einer subjekttheoretischen Überlegung reflektiert werden. Mit Rekurs auf poststrukturalistische Ansätze wird die Verschränkung von KI und Selbst noch genauer gefasst und zudem diskutiert, welche ethischen Fragen damit verbunden sind. KI entwickelt sich auch aus dieser Perspektive von einem Vertrauensobjekt hin zu einer Vertrauensneh-

merin. KI-Systeme sind nicht nur (und wohl auch nicht primär) passive, technologische Artefakte, sondern treten als handlungsfähige Akteurinnen auf, „as systems of networks and institutions" (Mohamed et al. 2020, 660).

Mit einer so verstandenen KI sind zudem implizite normative Vorstellungen verbunden. KI erzeugt nicht nur Identitätsmerkmale, sondern verändert auch grundlegend bisher etablierte Antworten auf normative Fragen wie z. B. die, was Menschsein ist. Wie Haraway (2006) am Beispiel des Cyborgs argumentiert, kann die technologische Entwicklung, von der KI einen wesentlichen Teil ausmacht, dazu führen, dass binäre Oppositionen, z. B. zwischen Mensch und Nicht-Mensch/Maschine unscharf oder sogar aufgebrochen werden. Das Selbstverständnis, Mensch zu sein, wird durch KI herausgefordert und teilweise transformiert (Irrgang 2005). Für die Frage nach Vertrauen in KI ist dies relevant, weil KI-Nutzer*innen einer solchen Normativität nicht (immer) als autonome Individuen begegnen, d. h. sie können sich den durch KI transportierten normativen Vorstellungen von Subjektivität und Selbst gegenüber nicht souverän verhalten.

Ein solches Verhältnis zwischen KI und Nutzer*in wird im Anschluss an ein poststrukturalistisches Verständnis von Subjektivität nachvollziehbar. Dies bedeutet, dass etablierte und als normal erlebte und verstandene Selbstverständnisse Produkt von Subjektivierungsmechanismen sind (z. B. Foucault 2022).[3] Die Art und Weise, wie Menschen sich selbst verstehen, ist Produkt von historisch gewachsenen institutionellen, pädagogischen oder wissenschaftlichen Diskursen. Formen von Subjektivität oder Identität sind deshalb immer schon, als normative Subjektivitätspositionen, in die subjektivierenden Gesellschaftsstrukturen eingeschrieben. Daran anschließend argumentiert bspw. Judith Butler für eine Ethik der Verletzbarkeit: Diese normative Heuristik impliziert, anzuerkennen, dass menschliche Selbstverständnisse und Handlungsfähigkeiten sozial verfasst und relational geformt sind (z. B. Butler 2005).

Dieses Konzept der Subjektivierung lässt sich nun auf KI-Systeme übertragen und damit die Wanderung des Vertrauens vom Menschen zur

3 Die Konzepte ‚Selbst', ‚Subjektivität', ‚Identität' und ‚Selbstverständnis' werden im Folgenden synonym verwendet. Selbstverständnisse (z. B. ‚ich bin ein Mensch') sind zentrale Inhalte von Subjektivität. Identität wird teilweise parallel zu Subjektivität verwendet, allerdings betonen verschiedene Autor*innen auch Unterschiede, z. B. wird Subjektivität häufiger individualistisch gefasst, während Identität stärker als kollektives Beschreibungsmerkmal verwendet wird, z. B. in nationalen Diskursen als nationale Identität (Zima 2017).

Maschine weiter erklären. Besonders deutlich tritt dies im Kontext der Digitalisierung alltäglicher Interaktionsformen und Praktiken zutage. KI kann, vermittelt bspw. über digitale Medien, zu einer „medienkulturellen Subjektivierung von Benutzer*innen" (Breljak and Mühlhoff 2019, 18) beitragen. Das Verhältnis von Nutzer*innen zu KI ist deshalb weniger eines zwischen autonomen Subjekten (mit klaren Erwartungshaltungen) und technologischen Objekten. Anja Breljak und Rainer Mühlhoff betonen stattdessen im Sinne eines poststrukturalistischen Verständnisses von Subjektivität: „An die Stelle des Subjekts, welches sich – so die klassische Vision – der technischen Artefakte rein instrumentell bedient, tritt das affektive Verhältnis zum vernetzten Gerät" (Breljak und Mühlhoff 2019, 17). Dieses Verhältnis zu (über digitale Medien vermittelten) KI-Systemen ist ein verletzbares, weil Subjekte durch dieses Verhältnis in ihren Handlungsmöglichkeiten massiv eingeschränkt oder gar aus sozialen Praktiken ausgeschlossen werden können. Mohamed et al. (2020) legen bspw. offen, wie Algorithmen koloniale Unterdrückung (z. B. durch predictive policing), Ausbeutung (z. B. Arbeitsbedingungen und wissenschaftliche KI-Experimente) und Enteignung (z. B. die Verfremdung und ‚Verwestlichung' indigener Daten) wiederholen und dadurch Subjekte als rassifiziert ansprechen. Insofern Nutzer*innen KI-Technologien in ihren alltäglichen Praktiken verwenden, übernehmen sie auch die durch die KI transportierten (z. B. rassifizierten) Subjektivitätsvorstellungen. KI produziert eine bestimme Weise des Denkens, Fühlens und Handelns (Mühlhoff 2023). Auch enthält KI das Potenzial, etablierte Selbstverständnisse des Menschseins herauszufordern, indem es Grenzziehungen zwischen Mensch und Maschine infragestellt (Prugger 2022). KI ist deshalb nicht nur eine passive Technologie, sondern tritt als subjektivierende Akteurin auf, die Subjektpositionen kreiert und reproduziert (z. B. hinsichtlich der rassistischen Einteilung von Menschen nach Hautfarben oder der anthropologischen Frage, was es bedeutet ein Mensch zu sein). Diese werden durch Nutzer*innen besetzt: Sie verstehen sich als das Subjekt, das durch die KI transportiert und angeboten wird. In der Verwendung von KI-gesteuerten Technologien macht sich eine Nutzer*in verletzbar, insofern ihr Selbst von KI produziert wird.

Das Selbst, das KI nutzt, wird dadurch also als verletzbares Subjekt angesprochen, das von KI subjektiviert werden kann. An dieser Stelle zeigt sich die Verbindung von KI, Verletzbarkeit und Vertrauen. Vertrauen und Verletzbarkeit stehen nach Baier in einem engen Verhältnis zueinander. „To trust is to let another think about and take action to protect and advance

something the truster cares about, to let the trusted care for what one cares about.“ (Baier 1992, 120). Vertrauen erfordert Verletzbarkeit. Im Falle von Vertrauen in KI ist das Objekt des Vertrauens, um welches sich Vertrauensgebende sorgen, ihr eigenes Selbst. Die Verletzbarkeit des eigenen Selbst zuzulassen, wird damit zur zentralen Voraussetzung von Vertrauen in KI.

Misstrauen in KI könnte im Anschluss an diese Überlegungen bspw. weniger als Resultat von enttäuschten oder irrationalen Erwartungshaltungen verstanden werden, sondern stärker als Ergebnis von Verletzungserfahrungen bzw. eines Schutzes der eignen Verletzbarkeit. „[B]etrayals of trust lead not just to loss of a particular entrusted good but to a lasting inability to partake of that sort of trust-dependent good. And if the trust-dependent goods are the most precious, then that is a severe disability.“ (Baier 1992, 133). Besteht die Gefahr, dass sich das Selbst durch die Nutzung von KI einer Form der Subjektivierung aussetzt, die verletzend, diskriminierend oder ausbeutend ist (wie bspw. durch eine spezifische Rassismen reproduzierende KI), kann dies zu einem allgemeinen Misstrauen in KI führen. Dieses Misstrauen zeigt sich dann nicht zwingendermaßen als eine rational gelesene Formulierung enttäuschter Erwartung, die ein autonomes und souveränes Subjekt zunächst in die KI-Nutzung hatte. Misstrauen in KI kann sich stattdessen als eine rigorose Weigerung zeigen, sich durch KI-Nutzung verletzbar zu machen, Kontrolle abzugeben und sich von KI subjektivieren zu lassen.

Einem solchen Verständnis nach fungiert KI nicht mehr als Vertrauensobjekt, demgegenüber sich rationale Erwartungen formulieren lassen. Aus der Perspektive von Verletzbarkeit als Voraussetzung von Vertrauen in KI wird deutlich, dass KI zur Akteurin der Subjektivierung wird. Für das Ziel eines vertrauensvollen Umgangs mit KI, tritt KI als Vertrauensnehmerin auf, der hinsichtlich eines Objekts des Vertrauens – dem eigenen Selbst – Vertrauen ausgesprochen wird.

5 Fazit und Ausblick

Die zentrale These des Beitrages ist, dass Vertrauen im Kontext mehr als Verlässlichkeit oder rationale Erwartbarkeit meint. Dabei wandert das Vertrauen angesichts von KI-Entwicklungen aus einem rein intersubjektiven Verständnis hin zu einem technizierbaren Konstrukt und es entstehen neue Vertrauensverhältnisse. Folglich handelt es sich um keine intersubjektiven Vertrauensverhältnisse im klassischen Sinn. Die Grenzen zwischen Vertrauensobjekt und Vertrauensnehmenden verschwimmen vielmehr, da die

Handlungsmacht algorithmischer Systeme auf der inhärenten Normativität der Digitalität sowie der technisch vermittelten Intentionalität (kollektive Dimension) von KI-basierten Systemen beruht. Deshalb greift auch das traditionelle Verständnis von Vertrauen als ein Sich-Verlassen auf die technische Verlässlichkeit von KI zu kurz. Die vorangegangenen Überlegungen verstehen sich als mögliche Argumentationslinien für eine Erweiterung des Vertrauensverständnisses im Kontext von KI.

Gerade vor dem Hintergrund aktueller gesellschaftlicher Diskussionen erscheint es dabei wichtig zu betonen, dass es eine grundlegende Differenz zwischen Menschen und Maschinen gibt. Auch komplexe und selbstlernende KI-Systeme sind menschengemachte technische Kulturprodukte, denen kein eigenes Bewusstsein oder eine eigene Form von Emotionalität zukommt, auch wenn KI-Systeme beides im Laufe der Zeit immer besser imitieren können. Es geht deshalb in den Diskussionen um KI nicht um Vertrauen als eine zwischenmenschliche Kategorie in einem herkömmlichen Sinne. Gleichzeitig können Mensch und KI-System auch nicht vollständig getrennt werden. Das bedeutet nicht, dass Maschinen wie Menschen sind, aber es wäre auch zu kurz gegriffen, Maschinen nur als passive Objekte zu verstehen, die lediglich technische Abläufe vollziehen.

Ein Verständnis von Vertrauen kann deshalb an den intersubjektiven Begriff anschließen, muss diesen jedoch grundlegend transformieren. Die drei Gedankenfiguren eröffnen Grundlinien hierfür. Zwei Überlegungen sollen abschließend als ein Ausblick skizziert werden, die mögliche Forschungsdesiderate im Anschluss an diese Transformationen formulieren.

Erstens gilt es für zukünftige Forschungen noch stärker als bislang die praxeologischen Kontexte der KI-Systeme zu rekonstruieren und zu problematisieren (Koska und Reder 2023). Vertrauen in KI ist wie gesehen kein neutrales, rationales oder rein individuelles Verhältnis zwischen zwei Polen (z. B. Individuum und Maschine), sondern manifestiert sich (oder wird gebrochen) in und durch soziale, ökonomische oder politische Praktiken. Für die Reflexion der Bildung von Vertrauen spielt der Blick auf diese Praktiken eine zentrale Rolle. In Anlehnung an die Methodik des philosophischen Pragmatismus, der die Rekonstruktion von Praktiken und den in diesen eingebetteten Erfahrungen und normativen Konflikten zum Ausgangspunkt der philosophischen Reflexion macht, ließen sich solche Praktiken noch klarer analysieren als dies bislang der Fall zu sein scheint (Koska und Reder 2023). Denn beim Umgang mit KI-Systemen geht es gegenwärtig meist weniger um das Vertrauen in die KI als solche, als vielmehr um das Vertrauen in den Umgang von Personen

und Institutionen mit diesen Systemen. In die Ermöglichung oder Förderung vertrauensvoller Praktiken fließen sowohl rationale als auch emotionale Komponenten ein. Personen und Institutionen schaffen einerseits Vertrauen, indem sie den Umgang mit KI transparent und rational erklärbar gestalten. Andererseits sind sie auch darauf angewiesen, eine emotionale Zustimmung zu ihrem Verhalten zu fördern. Dabei ist die Entwicklung konkreter Anwendungspraktiken auch mit überzogenen Anthropomorphismen konfrontiert, die zwar einerseits verständlich sind, andererseits aber nur allzu leicht den Unterschied zwischen Menschen und Maschinen zu nivellieren versuchen und damit eher Misstrauen schüren als Vertrauen fördern.

Für die philosophische Forschung geht es zweitens um eine stärkere Reflexion der Verbindung von Vertrauen und Verletzbarkeit – gerade angesichts der weitreichenden Folgen von KI in unterschiedlichen sozialen Feldern (Gutwald et al. 2021). In der vielfältigen Verschränkung von Prozessen der Subjektivierung des Selbst und KI-Prozessen zeigten die Überlegungen, dass das Selbst herausgefordert wird, sich als verletzbares zu verstehen. Vertrauen ist letztlich auf ein Zulassen von Verletzbarkeit angewiesen. Letzteres ist jedoch nicht für alle Menschen in gleicher Weise möglich, da ‚Sich-verletzbar-Machen' je nach sozialer Position unterschiedliche Folgen haben kann.

Ein Feld, das in diesem Zusammenhang für die praktische Philosophie gegenwärtig immer wichtiger wird, ist die Reflexion global ungleicher Verletzbarkeit. Diese wird beispielsweise im Kontext postkolonialer Debatten verstärkt diskutiert und es gilt für die Zukunft auszuloten, welche Konsequenzen sich daraus für den Umgang mit KI-Systemen ergeben. Wie bereits in Kapitel vier dargestellt, wurde aus postkolonialer Perspektive zuletzt verstärkt auf koloniale Kontinuitäten von KI-Praxiszusammenhängen hingewiesen. Kolonial bedingte Ausschlüsse durch KI-Medien zeigen sich an einer „data rationality" (Ricaurte 2019), die in der Sammlung, Speicherung, Verfügung, Verarbeitung und Verwendung von Daten neben anderen Ausschließungsheuristiken koloniale Muster reproduziert.

Für die Frage nach Vertrauen ergibt sich dadurch eine strukturell ungleiche Verteilung von Verletzbarkeit aufgrund jahrhundertelanger kolonialer Machtausübung. Dadurch ist ein weiteres ‚Sich-verletzbar-Machen' als Voraussetzung für Vertrauen für rassifizierte Subjekte mit höheren Gefahren verbunden. Dabei wird auch die skizzierte Gefahr der KI als Mechanismus der Subjektivierung verschärft, da KI-Praxiszusammenhänge koloniale Verletzungen häufig nicht nur ausblenden, sondern sogar verstärken können. Misstrauen in KI ist aus der Perspektive rassifizierter Subjekte somit eine erwartbare

und nachvollziehbare Haltung, die es ernst zunehmen gilt, wenn über Vertrauen in KI nachgedacht wird. Allerdings scheint es manchmal in den (philosophischen) Debatten nur um das Vertrauen der Mittel- und Oberschicht der Industrieländer in die KI zu gehen. Die vorangegangenen Überlegungen sind Anlass dafür, eine solche Engführung zurückzuweisen. Wenn Menschen sich mit Blick auf KI-gesteuerte Maßnahmen als rassifiziert und bestimmte KI-Systeme damit letztlich als gewalttätig erleben, führt dies automatisch zu einem massiven Vertrauensbruch. Dies unterstreicht die Dringlichkeit, gezielt Maßnahmen zu entwickeln und umzusetzen, die auf die vielfältigen Dimensionen struktureller Unterdrückung (wie Rassismus, Klassismus, Sexismus, Altersdiskriminierung, religiöse Intoleranz u.v.m.) reagieren. Solche Bemühungen sind entscheidend, um die Entwicklung und Anwendung von KI-Technologien inklusiver und gerechter zu gestalten und können einen wertvollen Beitrag zur Verringerung von Diskriminierung und Ungleichheit leisten.

In westlich-demokratischen Diskursen wird bei der Deutung des Politischen oftmals auf das Element der rationalen Begründbarkeit und Erwartbarkeit abgestellt. Dies gilt teilweise auch für die Deutung von KI. Die Überlegungen haben gezeigt, dass ein solches Verständnis zu kurz greift. Nur wenn es gelingt, in die vielfältigen gesellschaftlichen Praktiken, in denen KI-Systeme zum Einsatz kommen, auch die Frage nach der Verletzbarkeit praxeologisch und kontextsensitiv zu integrieren, ist ein vertrauensvoller Umgang mit KI möglich. Vor diesem Hintergrund können die Bemühungen der Europäischen Kommission – Kriterien für die Herstellung und Zertifizierung von ‚trustworthy AI' zu finden – als ein wichtiger Schritt in die richtige Richtung verstanden werden. Werkzeuge wie das ‚AI Ethics Canvas' (Götze 2021), die sich für eine praxis- und anwendungsorientierte Betrachtung sowie für die Integration von ethischen Produkt- und Systemanforderungen in der Designphase eignen, können bereits ergänzend zur Risikoklassifizierung des als ‚AI Act' bekannt gewordenen Vorschlags der Europäischen Kommission eingesetzt werden. Sie sind aber auch unabhängig von diesem nutzbar, um die Belange vulnerabler Gruppen beim Systemdesign zu berücksichtigen. Die systematische Anwendung und Weiterentwicklung von Methodiken, die auf mehr als Verlässlichkeit abzielen und auch die Verletzbarkeit als einen zentralen Indikator von Vertrauen in KI berücksichtigen, hängt letztlich immer von der Bereitschaft und dem Engagement der Herstellenden und Betreibenden von KI-Systemen ab. Diese freiwillige Übernahme von Verantwortung lässt sich von den Vertrauensnehmenden schon jetzt im Rahmen der ‚Corporate Digital Responsibility' (Filipović 2024) in der Unternehmenskultur verankern.

Literatur

Ananny, Mike, und Kate Crawford. 2016. „Seeing without knowing. Limitations of the transparency ideal and its application to algorithmic accountability". *New Media & Society* 33 (4): 1–17. https://doi.org/10.1177/1461444816676645.

Baier, Annette. 1986. „Trust and Antitrust". *Ethics* 96 (2): 231–60. https://doi.org/10.1086/292745.

———. 1992. „Trust." In *The Tanner Lectures on Human Values*, herausgegeben von Grethe Ballif Peterson, 109–174. Salt Lake City: University of Utah Press.

Becker, Heidrun. 2018. „Robotik in der Gesundheitsversorgung: Hoffnungen, Befürchtungen und Akzeptanz aus Sicht der Nutzerinnen und Nutzer". In *Pflegeroboter*, herausgegeben von Oliver Bendel, 229–48. Wiesbaden: Springer Fachmedien Wiesbaden.

Breljak, Anja und Rainer Mühlhoff. 2019. „Was ist Sozialtheorie der Digitalen Gesellschaft? Einleitung". In *Affekt Macht Netz. Auf dem Weg zu einer Sozialtheorie der Digitalen Gesellschaft,* herausgegeben von Rainer Mühlhoff, Anja Breljak, Jan Slaby, 1. Aufl., 7–34. Bielefeld: transcript. https://doi.org/10.25969/mediarep/13219.

Butler, Judith. 2005. *Gefährdetes Leben. Politische Essays*. Frankfurt am Main: Suhrkamp.

European Commission. 2021. *Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts COM (2021)* 206 final: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206.

Faulkner, Paul. 2015. „The attitude of trust is basic". *Analysis* 75 (3): 424–29. https://doi.org/10.1093/analys/anv037.

Filipović, Alexander. 2024. „Corporate Digital Responsibility". In *The Routledge Handbook of Responsibility*, herausgegeben von Maximilian Kiener, 419–430. New York, London: Routledge. https://doi.org/10.4324/9781003282242.

Foucault, Michel. 2022. *Überwachen und Strafen. Die Geburt des Gefängnisses 22. Aufl*. Frankfurt am Main: Suhrkamp.

Friedman, Batya, und Helen Nissenbaum. 1996. „Bias in computer systems". *ACM Trans. Inf. Syst*. 14 (3): 330–47. https://doi.org/10.1145/230538.230561.

Friedrich, Orsolya, Johanna Seifert und Sebastian Schleidgen. 2023. *Mensch-Maschine-Interaktion: Konzeptionelle, soziale und ethische Implikationen neuer Mensch-Technik-Verhältnisse*. 1. Aufl., Boston: BRILL.

Gloyna, Tanja. 2017. „Vertrauen". In *Historisches Wörterbuch der Philosophie online*, herausgegeben von Joachim Ritter, Karlfried Gründer, Gottfried Gabriel, Jann Holl, Hans Lenk und Matthias Maring. Basel: Schwabe Verlag. https://doi.org/10.24894/HWPh.7965.0692.

Götze, Trystan. 2021. „AI Ethics Canvas: Guidebook." Zugriff am 25. November 2023. https://www.trystangoetze.ca/_files/ugd/e90586_266a66787c2649f382a40f8032a6c559.pdf.

Grassegger, Hannes, und Mikael Krogerus. 2016. „Ich habe nur gezeigt, dass es die Bombe gibt." Zugriff am 8. Dezember 2016. https://www.dasmagazin.ch/2016/12/03/ich-habe-nur-gezeigt-dass-es-die-bombe-gibt.

Greenwald, Glenn. 2014. *No Place to Hide: Edward Snowden, the NSA, and the U.S. Surveillance State*. 1. Aufl. New York: Metropolitan Books Henry Holt.

Gutwald, Rebecca, Jennifer Burghardt, Maximilian Kraus, Michael Reder, Robert Lehmann, und Nicholas Müller. 2021. „Soziale Konflikte und Digitalisierung. Chancen und Risiken digitaler Technologien bei der Einschätzung von Kindeswohlgefährdungen". In: *EthikJournal* 7 (2): 1–20. Online verfügbar unter: https://www.ethikjournal.de/fileadmin/user_upload/ethikjournal/Texte_Ausgabe_2021_2/Gutwald_u.a._Ethikjournal_2.2021.pdf.

Haraway, Donna. 2006. „A Cyborg Manifesto: Science, Technology, and Socialist-Feminism in the Late 20th Century". In *The International Handbook of Virtual Learning Environments,* herausgegeben von Joel Weiss, Jason Nolan, Jeremy Hunsinger, Peter Trifonas, 117–58. Dordrecht: Springer.

Hartmann, Martin. 2020. *Vertrauen: Die unsichtbare Macht*. 1. Aufl. Frankfurt am Main: S. Fischer.

———. 2022. „Alle wollen Vertrauen – Niemand will vertrauen". LUKB Vorlesungsreihe, Universität Luzern, 22. Juni. Zugriff am 25. November 2023. https://www.youtube.com/watch?v=HtzEmk2xJTo.

Heesen, Jessica. 2018. „Wer entscheidet für uns? Big Data, intelligente Systeme und kluges Handeln". In *Mensch – Maschine. Ethische Sichtweisen auf ein Spannungsverhältnis,* herausgegeben von Petra Grimm und Oliver Zöllner, 1. Aufl., 47–57. Stuttgart: Franz Steiner Verlag.

Heise, Nele. 2016. „Algorithmen." In *Handbuch Medien- und Informationsethik*, herausgegeben von Jessica Heesen, 202–9. Stuttgart: J. B. Metzler.

IEEE, Institute of Electrical and Electronics Engineers (Hrsg.) 2021. "IEEE Standard Model Process for Addressing Ethical Concerns during System Design", in *IEEE Std 7000-2021*, 1–82, 15 Sept. 2021, https://doi.org/10.1109/IEEESTD.2021.9536679.

Irrgang, Bernhard. 2005. *Posthumanes Menschsein? Künstliche Intelligenz, Cyberspace, Roboter, Cyborgs und Designer-Menschen – Anthropologie des künstlichen Menschen im 21. Jahrhundert*. Stuttgart: Franz Steiner Verlag.

Jaume-Palasí, Lorena, und Matthias Spielkamp. 2017. *Ethik und algorithmische Prozesse zur Entscheidungsfindung oder -vorbereitung. Arbeitspapier 4.* Herausgegeben von Algorithm Watch. Online verfügbar unter: https://algorithmwatch.org/wp-content/uploads/2017/06/AlgorithmWatch_Arbeitspapier_4_Ethik_und_Algorithmen.pdf.

Jones, Karen. 1996. „Trust as an Affective Attitude". *Ethics* 107 (1): 4–25. https://doi.org/10.1086/233694.

Koska, Christopher. 2023. *Ethik der Algorithmen. Auf der Suche nach Zahlen und Werten*. Techno:Phil – Aktuelle Herausforderungen der Technikphilosophie (Band 6). Berlin, Heidelberg: J. B. Metzler. https://doi.org/10.1007/978-3-662-66795-8.

Koska, Christopher, und Michael Reder. 2023. „KI-gestützte Assistenz für moralische Konfliktsituationen. Zur Algorithmisierung im Handlungsfeld der Kindeswohlgefährdung". In: *Medien – Ethik – Digitalisierung. Aktuelle Herausforderungen*, herausgegeben von Petra Grimm und Oliver Zöllner, 1. Aufl., 19–36. Stuttgart: Franz Steiner Verlag.

Krämer, Sybille. 2020. „Kulturgeschichte der Digitalisierung." Vorlesungsreihe *Making sense of the digital society*, Alexander von Humboldt Institut für Internet und Gesellschaft. 13. Februar. https://www.youtube.com/watch?v=S8Mtjqil5_s

Manovich, Lev. 2005. *Black box – white cube*. Dt. Ausg., Internationaler Merve-Diskurs 263. Berlin: Merve-Verlag.

———. 2018. *Cultural Analytics*. Cambridge, MA, Berlin: MIT Press; Knowledge Unlatched. http://www.knowledgeunlatched.org/ku-select-2017-books.

Margreiter, Reinhard. 2003. „Medien/Philosophie: Ein Kippbild." In *Medienphilosophie: Beiträge zur Klärung eines Begriffs*, herausgegeben von Stefan Münker, Alexander Roesler und Mike Sandbothe, 150–71. Frankfurt am Main: S. Fischer.

Martini, Mario. 2019. *Blackbox Algorithmus – Grundfragen einer Regulierung Künstlicher Intelligenz*. Unter Mitarbeit von Michael Kolain und Jan Mysegades. Berlin: Springer.

McLeod, Carolyn. 1999. *Self-trust and reproductive autonomy*. Canadian theses = Thèses canadiennes. Ottawa: National Library of Canada = Bibliothèque nationale du Canada.

Meckel, Miriam. 2018. „Der Spion in meinem Kopf: Technik-Konzerne wollen in unser Gehirn vordringen. Gelingt ihnen das, steht die Freiheit unserer Gedanken auf dem Spiel". ZEIT online, 18. April. Zugriff am 31. August 2018. *https://www.zeit.de/2018/16/brainhacking-gehirn-kopf-konzerne-miriam-meckel*.

Michalski, Niels. 2019. *Normatives und rationales Vertrauen in Europa: Eine ländervergleichende Untersuchung gesellschaftlicher Vertrauensniveaus*. Wiesbaden: Springer Fachmedien Wiesbaden.

Mohamed, Shakir, Marie-Therese Png, und William Isaac. 2020. „Decolonial AI: Decolonial Theory as Sociotechnical Foresight in Artificial Intelligence". *Philosophy & Technology* 33 (4): 659–84. https://doi.org/10.1007/s13347-020-00405-8.

Mühlhoff, Rainer. 2023. *Die Macht der Daten. Warum Künstliche Intelligenz eine Frage der Ethik ist*. Paderborn: Vandenhoeck & Ruprecht GM.

Nordenbrock, Kay. 2022. „Jeff Bezos und Bill Gates investieren in Brain-Computer-Interface-Startup Synchron". Zugriff am 2. Januar 2023. https://t3n.de/news/jeff-bezos-bill-gates-brain-computer-interface-startup-synchron-1521463/?utm_source=rss&utm_medium=feed&utm_campaign=news.

Prugger, Julian. 2022. „Digitalisierung als Kritik von Subjektivität. Über Kontingenz und Politisierung von Menschsein in einer digitalen Welt". In *Menschsein in einer technisierten Welt: Interdisziplinäre Perspektiven auf den Menschen im Zeichen der digitalen Transformation*, herausgegeben von Eva-Maria Endres, Anna Puzio und Carolin Rutzmoser. Wiesbaden, 93–110. Wiesbaden: Springer VS. https://doi.org/10.1007/978-3-658-36220-1_7.

Reinhardt, Karoline. 2022. „Trust and trustworthiness in AI ethics". *AI Ethics* 3: 735–44. *https://*doi.org/10.1007/s43681-022-00200-5.

Ricaurte, Paola. 2019. „Data Epistemologies, The Coloniality of Power, and Resistance". *Television & New Media 20 (4), 350–365.* https://doi.org/10.1177/1527476419831640.

Rixecker, Kim. 2017. „Facebook will unsere Gedanken auf den Computer bringen". Zugriff am 3. September 2020. http://t3n.de/news/facebook-gedanken-computer-816174/.

Russell, Stuart J., und Peter Norvig. 2022. *Artificial Intelligence: A Modern Approach.* Fourth edition, global edition. Pearson series in artificial intelligence. Harlow: Pearson.

Sandbothe, Mike. 2003. „Der Vorrang der Medien vor der Philosophie". In *Medienphilosophie: Beiträge zur Klärung eines Begriffs*, herausgegeben von Stefan Münker, Alexander Roesler und Mike Sandbothe, 185–97. Frankfurt am Main: S. Fischer.

VDE, Verband der Elektrotechnik Elektronik Informationstechnik e. V., Hrsg. 2022. *VCIO based description of systems for AI trustworthiness characterization, VDE SPEC 90012 V1.0 (en).* Online verfügbar unter: https://www.vde.com/resource/blob/2242194/a24b13db01773747e6b7bba4ce20ea60/vcio-based-description-of-systems-for-ai-trustworthiness-characterisationvde-spec-90012-v1-0--en--data.pdf.

Wolf, Susanna, Jan-Fiete Schütte, Marc Hauer und Christopher Koska. „Vertrauen im Kontext. Messung und Operationalisierung". In *Künstliche Intelligenz und ethische Verantwortung*, herausgegeben von Michael Reder und Christopher Koska. Im Druck: transcript.

Zima, Peter V. 2017. *Theorie des Subjekts. Subjektivität und Identität zwischen Moderne und Postmoderne.* Stuttgart: Utb.