ERGEBNISBERICHT ZUM WORKSHOP

Vertrauenswürdige Künstliche Intelligenz - *Wie kann die praktische Umsetzung gelingen?*

Virtueller Expert:innen-Workshop des Forums Privatheit
zur Entwicklung und Gestaltung einer vertrauenswürdigen Künstlichen Intelligenz (KI) am 13.10.2021.

Einleitung

Aus der Entwicklung und dem Einsatz von KI-basierten Lösungen ergeben sich neuartige Chancen ebenso wie neue Anforderungen für das verantwortliche Handeln im Umgang mit dieser Technologie. Wo stehen wir bei der praktischen Umsetzung einer vertrauenswürdigen KI heute, welche Umsetzungsdefizite bestehen und welchen Beitrag kann die partizipative Technikgestaltung leisten? Diese und weitere Fragen wurden im Rahmen des interaktiven Workshops: „Vertrauenswürdige KI: Wie kann die praktische Umsetzung gelingen?" am Mittwoch, den 13. Oktober 2021 virtuell diskutiert. Der Workshop wurde im Rahmen des BMBF-Projekts PRIDS[1] organisiert und bestand aus den drei thematischen Blöcken: Umsetzung, Strategien und Partizipation. Die thematischen Blöcke wurde durch Impulsvorträge von folgenden Speaker:innen eröffnet:

- Thilo Hagendorff (Uni Tübingen)
- Dr. Christian Winkler (Datanizing)
- Lajla Fetic (Bertelsmann Stiftung)
- Dr. Christoph Peylo (Bosch)
- Dr. Andreas Bischof (Uni Chemnitz)

Insgesamt nahmen 25 KI-Expert:innen aus den Bereichen Wirtschaft und Wissenschaft an dem Workshop teil. Die unterschiedlichen fachlichen Hintergründe ermöglichten eine intensive interdisziplinäre Auseinandersetzung mit dem Thema vertrauenswürdige KI. Der Bericht gibt einen Überblick über den Verlauf, die Inhalte sowie die Diskussionsergebnisse der Veranstaltung und schließt mit einer Darstellung von Herausforderungen und Lösungsvorschlägen für die praktische Umsetzung einer vertrauenswürdigen KI ab.

Inhalt

---

1 PRIDS: Privatheit, Demokratie und Selbstbestimmung im Zeitalter von Künstlicher Intelligenz und Globalisierung, www.forschung-it-sicherheit-kommunikationssysteme.de/projekte/prids

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

1. Zusammenfassung

Expert:innen aus Wissenschaft, Wirtschaft und Technikfolgenabschätzung (TA) präsentierten im Rahmen des Workshops eine Reihe von Ansätzen, wie ein verantwortungsbewusster Umgang mit KI in der Praxis gelingen kann und berichteten über Hürden, die es bei praktischen Anwendung ethischer Prinzipien zu überwinden gilt. Die gesetzten thematischen Schwerpunkte (Umsetzung, Strategien, Partizipation) wurden dabei nicht losgelöst voneinander abghandelt, sondern wurden durch die Diskussionen verzahnt und miteinander in Verbindung gebracht. Die Teilnehmer:innen des Workshops waren sich einig im Hinblick auf die Relevanz des Konzepts der vertrauenswürdigen KI, das 2019 von der High-Level-Expert-Group der EU entwickelt wurde und das aus sieben Anforderungen besteht, die vom Vorrang menschlichen Handelns vor maschinellen Entscheidungen über den Datenschutz und die Nichtdiskriminierung bis hin zur Rechenschaftspflicht reichen (High-Level Expert Group on Artificial Intelligence 2019). Als zentrale Herausforderungen wurden die Abstraktheit von Ethik-Guidelines, innerorganisationale Implementierungshürden, fehlende inhaltliche Spezifizierung des Konzepts, der Stand des KI-Ethik-Diskurses und die Ermöglichung von menschenzentrierter sowie partizipativer Technikgestaltung diskutiert. In den Impulsvorträge und in den Expert:innen-Beiträgen wurde herausgearbeitet, wie die Umsetzung von vertrauenswürdiger KI trotz dieser komplexen Herausforderungen gelingen kann.

2. Workshopinhalte und Diskussionsergebnisse

2.1 Umsetzung: Wo stehen wir bei der praktischen Umsetzung?

Impulsgeber
Dr. Thilo Hagendorff (Uni Tübingen)
Dr. Christian Winkler (Datanizing)

Dr. Thilo Hagendorff ist Experte für angewandte Ethik, insbesondere Technik- und KI-Ethik an der Universität Tübingen. Der Wissenschaftler eröffnete den Workshop mit einem Impuls zu den sechs Hürden bei der praktischen Umsetzung einer vertrauenswürdigen KI, die im Folgenden zusammenfassend dargestellt werden:

1. **Abstraktionsniveau:** Bestehende Richtlinien sind zu abstrakt und oftmals "technisch uninformiert", was bedeutet, dass sie nicht die gesamte Bandbreite der Methoden des maschinellen Lernens berücksichtigen, sondern ausschließlich auf Deep Learning oder überwachtes Lernen eingehen.
2. **KI Ethik ist nicht bindend:** Eine Abweichung von den Richtlinien hat für Unternehmen keine Konsequenzen.
3. **Wirtschaftsinteressen** konterkarieren ethische Normen.
4. **KI-Ethik als Marketingstrategie**: Technik wird ohne Berücksichtigung der Ethik-Prinzipien entwickelt, diese werden jedoch öffentlich kommuniziert ("ethic washing").
5. **Testing & Auditing**: Hohe Schwierigkeit die KI-Systeme zu auditieren.
6. "**bounded ethicality**": Ethisches Handeln ist aufgrund von internem und externem Druck oft eingeschränkt: Wir wissen, was richtig ist, handeln aber nicht danach.

Ferner beschrieb der Ethik-Forscher verschiedene Entwicklungsphasen einer vertrauenswürdigen KI aus wissenschaftlicher Perspektive: In der ersten Phase wurden seitens vieler Unternehmen prinzipienbasierte KI-Ethik-Richtlinien veröffentlicht. Diese bestehende Abstraktheit der Richtlinien wurde in einer zweiten Phase durch Expert:innen kritisiert ("practical turn"). Weiter beschrieb Hagendorff die dritte und aktuelle Phase, in der professionelle KI-Ethiker:innen "Ethics as a Service" propagieren und sich beim Auditing von Organisationen einbringen sowie seitens der Unternehmen weiterhin ein prinzipienorientierter Ansatz verfolgt wird. In der sich heute bereits abzeichnenden vierten Phase wird KI in einen Rechtsrahmen überführt (Artificial Intelligence Act). Für den wissenschaftlichen Ethik-Diskurs prognostizierte Hagendorff die Abweichung von Prinzipien - hin zu KI-Tugenden (AI Virtues), einer Fokussierung auf motivationale Settings und Persönlichkeitsdisposition.

Der AI Virtues Ansatz nach Hagendorff (2021) setzt auf der Ebene von Entwickler:innen an und umfasst die vier Tugenden *justice, honesty, responsibility* und *care.* Diese werden als Charakterdispositionen konzeptualisiert, welche die Praxis der Technikentwicklung bestimmen sollen[2]. Der AI-Virtues_Ansatz setzt auf eine Sensibilisierung für den ethischen Umgang mit der Technologie in unterschiedlichen Kontexten und Situationen, ohne dabei die technische Komplexität von KI zu vernachlässigen. Um den KI-Ethik-Diskurs praxisnäher zu gestalten, sollte nach Hagendorff zusätzlich der Fokus von reinen Ethikprinzipien auf prozessorientierte Ansätze gelegt werden, die anwendungsspezifische Regeln für den Umgang mit der Technologie ermöglichen.

Dr. Christian Winkler vom Startup datanizing stellte dem wissenschaftlichen Diskurs eine Anwendungsperspektive zur praktischen Umsetzung einer vertrauenswürdigen KI gegenüber. datanizing hat ein Text Analytics-Tool entwickelt, mit dem „Personas“, d.h. abstrahierte Kundenprofile,die in der Marktforschung genutzt werden, automatisch aus webbasierten Inhalten zu erstellen. Zunächst erläuterte der Unternehmer, dass die technologischen Entwicklungen rund um KI verschiedene Abwandlungen lernender Algorithmen hervorgebracht haben, die auf jeweils verschiedenen mathematischen Verfahren beruhen. Das Maschinelle Lernen (Machine Learning) beschrieb Winkler als Teilmenge der KI, bei dem Algorithmen aus Daten „lernen“. Wiederum eine Teilmenge vom Maschinellen Lernen sei das Deep Learning, bei dem verschiedene neuronale Netze verwendet werden.

Weiter gab Herr Winkler praxisnahe Einblicke in die Verwendung von KI in seinem Startup. datanizing verwende Algorithmen, um Muster in Daten zu erkennen, aber auch Vorhersagen zu treffen, um beispielsweise Änderungen und Verbesserungen in Geschäftsabläufen zu erzielen (Advanced Analytics). Nach Winkler wird die "fortgeschrittene Analytik" eingesetzt, um Datensätze zu strukturieren und aufzubereiten, allerdings müssen die Mitarbeiter:innen in der Lage sein, diese zu interpretieren. Ein solches KI-System würde nicht eingesetzt werden, um "automatisierte Entscheidungen" zu treffen, sondern um Routineaufgaben (bspw. Clusterung von Webeinträgen) zu übernehmen und Mitarbeiter:innen bei der Datenaggregation zu unterstützen. Hinsichtlich der Umsetzung ethischer Richtlinien stellte Herr Winkler zwei

[2] Hagendorff, T. (2021). AI virtues--The missing link in putting AI ethics into practice. arXiv preprint arXiv:2011.12750.

praxisbezogene Problemfelder dar: Zum einen würde der Fokus von Unternehmen darauf liegen, profitabel zu arbeiten und Produkte anzubieten, die Kundenerwartungen und -wünschen entsprechen. Zum anderen könnten algorithmenbasierte Auswertungen auf Basis von "voreingenommenen" Daten erfolgen. Auch die Verwechslung von Korrelation und Kausalität führe oftmals zu fehlerhaften Analysen.

Diskussionsergebnisse

- **Praxistauglichkeit von Ethik-Guidelines verbessern**

  Seitens der Teilnehmer:innen wurde festgehalten, dass Gestaltung, Entwicklung und Einsatz algorithmenbasierter Systeme derzeit weitgehend in einem regulatorischen und ethischen Vakuum erfolgt, sodass sich die Frage nach normativen Kriterien stellt, die als Maßstab für eine vertrauenswürdige Technologieentwicklung gelten können. Im Workshop wurden die mangelnde Praxistauglichkeit bestehender Ethik-Guidelines und ihre nicht generalisierbare Anwendbarkeit als Hemmnisse bei der Etablierung eines verantwortungsvollen Umgangs mit KI identifiziert. Als Beispiel hierfür wurde angeführt, dass die Regel "Man soll nicht diskriminieren" zu unkonkret sei, sodass Anwender:innen und Entwickler:innen keine anwendungsnahe Umsetzung dieser Regel vornehmen könnten. Es fehle an harmonisierten Standards die beispielsweise vorgeben, wann ein Diskriminierungsrisiko vorliege. Nur so könne man den Entwickler:innen eine ausreichende Rechtssicherheit geben, ob im Rahmen der Entscheidungsprozesse eines KI-Systems eine legale oder illegale Differenzierung vorliege.

- **Unternehmensethik und -kultur im Umgang mit KI aufbauen**

  In den Unternehmen muss laut der Expert:innen das Thema Ethik verstärkt in den Fokus interner Prozesse und Problemlösungslogiken rücken. Auf Ebene der Unternehmensorganisation müssen Prozesse und Strukturen für einen verantwortungsvollen Umgang mit KI in der Praxis etabliert werden. Als Beispiel wurde angeregt zu konzeptualisieren, wie die KI-Tugenden (siehe Hagendorff, 2021) praxisnah übersetzt und in interne Strukturen und Prozesse umgesetzt werden können. Hierbei spielt laut Expert:innen auch der Aspekt der internen Unternehmenskommunikation eine große Rolle. In den Unternehmen sollte Raum für Dialoge geschaffen werden, die einen internen Austausch zu Best Practices oder Problemen zwischen Entwickler:innen dienen. Dabei müsse herausgestellt werden, dass ein verantwortungsvoller Umgang mit der Technologie für den organisationalen Erfolg förderlich ist. Hierfür ist es zunächst erforderlich, die organisatorischen Voraussetzungen zu schaffen, damit die Ziele, die mit der neuen Technologie realisiert werden könnten, auch eine Chance auf Verwirklichung hätten. Dies sei allerdings sehr anspruchsvoll und setze auch voraus, dass wir unsere Wirtschaft transformieren, u.a. in die Richtung, die gegenwärtig unter dem Begriff „New Work“ diskutiert werde.

2.2 Strategien: Welche Gründe gibt es für das Umsetzungsdefizit?

Impulsgeber:innen
Lajla Fetic (Bertelsmann Stiftung)
Dr. Christoph Peylo (Bosch)

Lajla Fetic leitet das Projekt „Ethik der Algorithmen" der Bertelsmann Stiftung und beschäftigt sich hier mit ethischen Aspekten zur Entwicklung und Gestaltung von KI. Sie argumentierte, dass es für die Umsetzung von vertrauenswürdiger KI eines ganzen Werkzeugkastens von Checklisten, Leitfäden und Tools bedarf. In einem Gemeinschaftsprojekt war Frau Fetic 2019 an der Entwicklung der Algo.Rules beteiligt, eines Katalogs von formalen Regeln für die verantwortungsvolle Gestaltung algorithmenbasierter Systeme aus Perspektive von KI-Entwickler:innen (s. Abb. 1).

Abbildung 1: Übersicht Algo.Rules. Regeln für die Gestaltung algorithmischer Systeme

Quelle: Präsentation Lajla Fetic, siehe auch https://algorules.org/fileadmin/files/alg/Algo.Rules_DE_2.pdf

Frau Fetic erläuterte, dass die Algo.Rules 2019 einen Nerv getroffen hätten und auf großes öffentliches Interesse gestoßen seien. Dem gegenüber stehe jedoch, dass das Regelwerk bislang nicht flächendeckend von Unternehmen umgesetzt werde. Um die Implementierung anzuregen, entwickelte das Algo.Rules-Team daher einen Praxisleitfaden, der Orientierungshilfen für Entwickler:innen und ihre Führungskräfte geben und die Umsetzung der Algo.Rules anleiten soll. Damit die praktische Umsetzung einer vertrauenswürdigen KI gelingen kann, müsse nach Frau Fetic deutlicher herausgestellt werden, dass Guidelines, Checklisten und Tools darauf zielen, KI-Anwendungen „gut" und „sinnstiftend" zu implementieren. Das könne ihrer Meinung nach zu einer größeren Akzeptanz der Guidelines in Unternehmen führen. Nach der Expertin müsse es darum gehen, die Entwicklungsprozesse der Technologie besser zu gestalten und nicht darum, ethische Dimensionen im Nachhinein abzuhaken. Darüber hinaus plädierte sie dafür, eine Schär-

fung des Begriffsverständnisses vorzunehmen. Das Konzept der vertrauenswürdigen KI sei auch nicht zuletzt wegen seiner Mehrdimensionalität bisher nicht einheitlich definiert und werde deshalb je nach Anwendungsbereich unterschiedlich verwendet.

Dr. Christian Peylo leitet das Bosch Center for Artificial Intelligence und argumentierte im Rahmen seines Impulses, dass nicht nur das Verhalten der Menschen bei der Entwicklung von KI vertrauenswürdig sein müsse. Vielmehr sollten seiner Meinung nach KI-Systeme entwickelt werden, die vertrauenswürdig seien. Hierunter versteht Herr Peylo, dass das "Verhalten" der Systeme transparent und nachvollziehbar ist, sodass die Interaktion zwischen Mensch und KI-System möglich wird. Zugleich unterstrich Peylo die Notwendigkeit nach Verhaltenskodizes. Für ihn ist KI eine Technologie, die die Grundprinzipien des menschlichen Selbstverständnisses und der menschlichen Gesellschaft in Frage stellt. Um Vertrauen sowohl in die Technologie selbst als auch in die Kompetenz und das Risikobewusstsein zu schaffen, haben viele Unternehmen damit begonnen, Verhaltenskodizes zu veröffentlichen, die ihre Standards für den Einsatz von KI sichtbar machen. Da derzeitigen Produkthaftungsgesetze und Sicherheitsvorschriften nicht auf dynamische und adaptive Systeme ausgelegt sind, müssen sich laut Peylo Unternehmen die Frage stellen, wie die Einhaltung von Sicherheitsstandards mit „lernenden“ Systemen gewährleistet werden könne. Darüber hinaus prognostiziert der Experte, dass selbstverpflichtende Regeln und ethische Standards die Notwendigkeit einer Zertifizierung durch staatliche Behörden relativieren könne. Bosch entwickelte mittels interner und externe Expertise einen Kodex für eine verantwortungsvolle Entwicklung von KI. Hierbei unterscheidet der Kodex drei Szenarien der systemischen Mensch-Maschine-Interaktion[3]:

1. Der Mensch wird bei der Klassifikation von Gegenständen oder Lebewesen unterstützt (Human-in-command).
2. Bei sogenannten Human in the Loop Systemen treffen KI-Systeme selbst Entscheidungen, die der Mensch übersteuern kann (Human-in-the-loop).
3. Experten legen während der Entwicklung des KI-Systems/Produkts bestimmte Parameter als Grundlage für die Entscheidung der KI fest, in die Entscheidung selbst können sie nicht eingreifen (Human-on-the-loop).

Herr Peylo erläuterte, dass die Operationalisierung und Implementierung der Regeln für die verantwortungsvolle Entwicklung von KI-Systemen bei Bosch in einem Zweischritt-Verfahren sichergestellt werden soll: Zum einen bestehen Qualitätskriterien, die an festgelegten Meilensteinen bei der Entwicklung abgeprüft werden. Diese Kriterien werden auf existierende Entwicklungsverfahren "gemappt" und anschließend mit den bestehenden KI-Kodex-Guidelines referenziert. Darüber hinaus beteiligt sich Bosch im Digital Trust Forum. Die Initiative hat das Ziel, vertrauenswürdige digitale Lösungen für vernetzte, intelligente, physische Produkte zu ermöglichen.

[3] siehe ausführlich https://www.bosch-presse.de/pressportal/de/de/ki-kodex-bosch-gibt-sich-leitlinien-fuer-den-umgang-mit-kuenstlicher-intelligenz-208384.html

Diskussionsergebnisse

- **Inhaltliche Spezifizierung von vertrauenswürdiger KI, Vertrauen und Verantwortung vornehmen**

  Zunächst sind laut den Expert:innen bessere definitorische Grundlagen erforderlich, um durch eine weitere inhaltliche Spezifizierung der Konzepte Vertrauen und Verantwortung im Zusammenhang mit KI zu entwickeln und die zielgerichtete Diskussion der verschiedenen Stakeholder zu ermöglichen. In der Diskussion wurde hervorgehoben, dass ein relationales Begriffsverständnis von Vertrauen (*wer vertraut wem, in welcher Hinsicht, warum, und unter welchen Umständen*?) die definitorische Tiefenschärfe verbessern könne. Geht man nämlich von einem differenzierteren Begriffsverständnis aus, so stellen sich Vertrauen und Verantwortung nicht mehr nur als normative Orientierungsmaßstäbe dar, sondern als Handlungen voraussetzende, intentionale Selbst-Positionierung. Deshalb sollte bei der praktischen Umsetzung von KI explizit die Frage nach dem Ver- oder Misstrauen gestellt werden, das einer KI-Lösung entgegengebracht werde, denn Vertrauen werde in jeweils spezifischen sozialen Handlungszusammenhängen operationalisiert.

- **KI-Ethik-Diskurs anwendungsnah führen**

  Ferner wurde festgestellt, dass der wissenschaftliche KI-Ethik-Diskurs Ansätze, Begriffe und Konzepte hervorgebracht habe, die umfassend und stringent seien, die jedoch nur schwer in der Praxis anzuwenden seien. Ethische Reflexionen sollten laut den Teilnehmer:innen zukünftig verstärkt entlang von anwendungsnahen Belangen bei der Entwicklung und dem Einsatz einer vertrauenswürdigen KI erfolgen.

2.3 Partizipation: Welchen Beitrag kann die partizipative Technikgestaltung leisten?

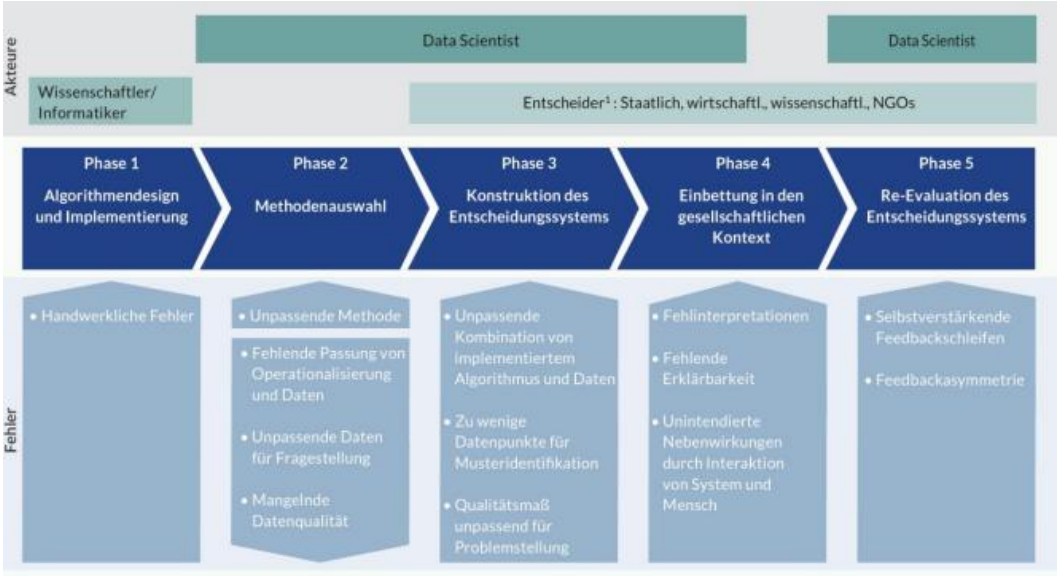
Impulsgeber
Dr. Andreas Bischof (Uni Chemnitz)

Dr. Andreas Bischof forscht an der Technischen Universität Chemnitz u.a. zu partizipativen Methoden der Technikentwicklung und leitet eine Forschungsgruppe des vom BMBF geförderten Projekts Miteinander. Ziel des Projekts ist es, Co-Design-Werkzeuge, Methoden und Dialogformate zu kreieren, die eine gesamtgesellschaftliche Teilhabe an sozio-technischen Entwicklung fördern.

In seinem Vortrag erläuterte Herr Bischof den Beitrag der partizipativer Technikgestaltung bei der Entwicklung einer vertrauenswürdigen KI. Partizipation ist für Bischof ein inhärent normatives Konzept, es hat politische Implikationen und kann anhand verschiedener methodischer Zugänge unterschiedlich beleuchtet werden. Akteure können in einem partizipativen Prozessen zwei Rollen einnehmen: Zum einen

könnten sie als "Subjekte" gesehen werden, die in partizipativen Prozessen reaktiv agieren (reactive informers). Zum anderen könnten Akteure aktiv gestalten und Prozesse mitsteuern (active co-creators)[4]. Herr Bischof machte deutlich, dass es nicht die eine Partizipation gibt. Stattdessen müssten je nach Aufgabe, Kontext, Beteiligten und Projektstand unterschiedliche Formen von Teilhabe ermöglicht werden. In Bezug auf die Entwicklung von algorithmischen Entscheidungsprozessen plädierte Bischof dafür, Partizipation in jeder Projektphase mitzudenken (s. Abb. 2). Aber auch die grundlegende Frage, ob KI in einem Anwendungsszenario eingesetzt werden soll, wird seiner Meinung nach selten partizipativ beantwortet. Partizipative Ansätze sollten nach Bischof verstärkt auch zu Beginn und gegen Ende des Technologieentwicklungsprozesses eingebracht werden. Dies verdeutlichte er anhand des Schaubildes zum Entwicklungs- und Einbettungsprozess von algorithmischen Entscheidungssystemen nach Zweig et al. 2018 (s. Abb.2)[5]:

Abbildung 2: Entwicklungs- und Einbettungsprozess von algorithmischen Entscheidungssystemen



Quelle: Zweig et al. 2018

In dem letzten Abschnitt seines Vortrags ging Bischof auf die Frage ein, wie Partizipation in Entwicklungsprozesse eingebracht werden kann. Hierfür gab er Einblicke in ein BMBF-Projekt, welches der Frage nachging, wie Maschinelles Lernen über Daten aus dem Smart Home Bereich zu Erkenntnissen im Hinblick auf die Sicherheit von Wohnen im Alter führen kann. Nach Ablauf des Beobachtungszeitraums wurden die generierten Datenkurven (und Erkenntnisse hieraus) in Dialoggruppen, bestehend aus Bewohner:innen und Entwickler:innen besprochen. Anhand der Offenheit des methodischen Designs wurde es möglich, die Informationsasymmetrie zwischen Entwickler:innen und Nutzer:innen aufzuheben. Darüber hinaus konnten Datenpraktiken von Laien beobachtet werden, sodass das Verständnis seitens der Techniker:innen für Interpretationen, Anwendungen und auch für alltagsweltlich verankerte Relevanzstrukturen wie Privacy und Vertrauen stieg.

[4] siehe auch Sanders, E. B. N. & Stappers, P. J. (2008). Co-creation and the new landscapes of design. *Co-design*, *4*(1), 5-18.

[5] Zweig, K. A., Fischer, S. & Lischka, K. (2018). Wo Maschinen irren können. Fehlerquellen und Verantwortlichkeiten in Prozessen algorithmischer Entscheidungsfindung. Bertelsmann Stiftung: Gütersloh.

Diskussionsergebnisse

- **Fokus auf eine menschenzentrierte Technikgestaltung**

  Unternehmen, die KI entwickeln und nutzen, stehen vor der Herausforderung, die Systeme effizient einzusetzen, aber auch den externen, also politischen und gesellschaftlichen Anforderungen gerecht zu werden. Dabei stehen zunächst die internen Logiken und Zielsetzungen der Unternehmen im Vordergrund. Folglich diskutierten die Expertinnen und Experten, wie ein verantwortungsvoller Umgang mit der Technologie in interne Strukturen und die Arbeitsorganisation übersetzt werden kann. Ein besonderer Fokus müsse hierbei auf den Personen im Unternehmen liegen, die mittel- oder unmittelbar mit Technologie in Berührung kommen. Zu einer holistischen Operationalisierung von Vertrauen in KI gehöre, dass seitens der Unternehmen Maßnahmen ergriffen werden, die den Menschen in den Mittelpunkt des Technologieeinsatzes stellen. Ein Beispiel hierfür sei der Human in the Loop-Ansatz, bei dem der Mensch die Ergebnisse von KI-Entscheidungssystemen jederzeit übersteuern, beeinflussen und verändern könne. Ein ethisches und verantwortungsbewusstes Verhalten in Bezug auf den Umgang mit der Technologie müsse von Anfang an in interne und externe Verarbeitungsprozesse angekoppelt werden.

- **Partizipation entlang des gesamten Technologieentwicklungsprozesses ermöglichen**

  Die Expert:innen diskutierten, dass eine verantwortungsvolle Entwicklung sowie der Einsatz von KI als partizipativer Gestaltungsprozess aufgefasst und erarbeitet werden sollte. Dieser Prozess ziele auf eine sozialverantwortliche Technikentwicklung und umfasse die Etablierung ethischer Rahmenbedingungen. Angebote zur aktiven Mitgestaltung der Technologieentwicklung sollten seitens der Unternehmen gefördert und ermöglicht werden, sodass verschiedene Akteure und nicht zuletzt die Anwender:innen in den Technologiegestaltungsprozess miteinbezogen werden. Dabei sei zu berücksichtigen, dass "die Gesellschaft" nicht partizipieren könne. Vielmehr diskutierten die Expert:innen, dass Advokaten und Akteure gefunden werden müssten, die auf der Basis gesamtgesellschaftlich geteilter Grundsätze argumentierten und sich aktiv in KI-Entwicklungsprozesse einbringen. Als Beispiel hierfür wurden zivilgesellschaftliche Organisationen angeführt.

3. Fazit: Herausforderungen und Lösungsansätze

Als Fazit des Workshops kann festgehalten werden, dass ethische Reflektionen zur Technologieentwicklung in Bezug auf die Gestaltung von vertrauenswürdiger KI zunächst auf der Ebene der Unternehmen ansetzen müssen. Unternehmen stehen vor der Herausforderung, KI nicht nur zur Umsatz- und Effizienzsteigerung einzusetzen, sondern auch die gesellschaftspolitischen Implikationen mitzudenken und sich im

hinblick auf die Forderung nach einen verantwortungsvollen Umgang mit der Technologie zu positionieren. Der Workshop ergab Herausforderungen aber auch praxistaugliche Anhaltspunkte und Lösungsansätze dafür, wie die praktische Umsetzung von verantwortungsvollem Unternehmenshandeln bei der Entwicklung und dem Einsatz von KI gelingen kann. Im Folgenden werden die Inhalte der Impulsvorträge sowie die Ergebnisse aus den Diskussionen noch einmal zusammenfassend anhand folgender Tabelle dargestellt:

Tabelle: Herausforderungen und Lösungsansätze für die praktische Umsetzung von vertrauenswürdiger KI

Herausforderungen	**Lösungsansätze**
Abstraktheit von Ethik-Guidelines	Praxistauglichkeit von Ethik-Guidelines verbessern
Innerorganisationale Implementierungshürden	Schaffung organisationaler Voraussetzungen in Entwicklung und Management für den vertrauenswürdigen Umgang mit KI
Fehlende definitorische Grundlagen und inhaltliche Spezifizierung von vertrauenswürdiger KI	Inhaltliche Spezifizierung von vertrauenswürdiger KI, Vertrauen und Verantwortung vornehmen
Stand des KI-Ethik-Diskurses	Ethische Reflexionen sollten zukünftig verstärkt entlang von anwendungsnahen Belangen erfolgen
KI-Systeme menschenzentriert gestalten	Arbeitsprozesse nach dem Human-in-the-Loop-Prinzip ausrichten
Partizipative Ansätze verfolgen	Partizipation von Akteuren und Advokaten entlang des gesamten Technologieentwicklungsprozesses ermöglichen

Quelle: Eigene Darstellung

Der Ergebnisbericht ist im Kontext des PRIDS-Projekts (BMBF) entstanden und steht außerdem im Zusammenhang mit Arbeiten der KI-Gruppe im Fraunhofer ISI. Ansprechpartner für das Thema vertrauenswürdige KI im Fraunhofer ISI sind Dr. Bernd Beckert, Dr. Michael Friedewald, Murat Karaboga, Greta Runge und Dirk Kuhlmann.
