RESEARCH ARTICLE

Vertrauenswürdige künstliche Intelligenz

Ausgewählte Praxisprojekte und Gründe für das Umsetzungsdefizit

Bernd Beckert, *Fraunhofer-Institut für System- und Innovationsforschung ISI, Breslauer Straße 48, 76139 Karlsruhe, DE*
(bernd.beckert@isi.fraunhofer.de) *0000-0003-0157-9096*

Zusammenfassung • Während es inzwischen eine ganze Reihe praktischer Leitfäden für die Implementierung des Konzepts der vertrauenswürdigen künstlichen Intelligenz (KI) gibt, fehlt es an konkreten Beispielen und Projekten für Umsetzungen, anhand derer sich Probleme und Erfolgsstrategien der Akteur:innen vor Ort aufzeigen ließen. Dieser Beitrag stellt ausgewählte Umsetzungsprojekte vor. Durchweg zeigt sich dabei ein noch geringer Grad an Konkretisierung. Deshalb wird anschließend nach den Gründen für das Umsetzungsdefizit gefragt. Drei Erklärungen kommen infrage: Time-to-Market-Überlegungen aufseiten der Unternehmen, Unklarheit darüber, welche Aspekte des Konzepts der vertrauenswürdigen KI bei welchen Anwendungen überhaupt relevant sind sowie die Tatsache, dass die Umsetzung von KI-Projekten komplexer ist als die Umsetzung ‚normaler' Software-Projekte und deshalb spezifische Vorkehrungen notwendig sind.

Trustworthy artificial intelligence. *Selected practical projects and reasons for the implementation deficit*

Abstract • *While there are now a number of practical guides for implementing the concept of trustworthy artificial intelligence (AI), there is a lack of concrete examples and projects for implementations that could be used to highlight problems and success strategies of actors in the field. This paper presents selected implementation projects showing that the degree of concretization is still low throughout. Therefore, the reasons for the implementation deficit are then explored. There are three possible explanations: time-to-market considerations on the part of the companies, lack of clarity about which aspects of the concept of trustworthy AI are relevant at all for which applications, and the fact that the implementation of AI projects is more complex than the implementation of 'normal' software projects and thus requires specific arrangements.*

This is an article distributed under the terms of the Creative Commons Attribution License CCBY 4.0 (https://creativecommons.org/licenses/by/4.0/)
https://doi.org/10.14512/tatup.30.3.17
Received: Jun. 13, 2021; revised version accepted: Oct. 20, 2021;
published online: Dec. 20, 2021 (peer review)

Keywords • *trustworthy AI, human centered AI, best practices, implementation gap, AI regulation*

Das Konzept der vertrauenswürdigen künstlichen Intelligenz

Unter vertrauenswürdiger künstlicher Intelligenz versteht man automatisierte Erkennungs-, Vorschlags- und Entscheidungssysteme, die den Anforderungen von Transparenz, Verantwortlichkeit, Privatheit, Diskriminierungsfreiheit und Zuverlässigkeit gerecht werden. Hintergrund ist die Tatsache, dass auf künstlicher Intelligenz (KI) basierende Systeme gesellschaftliche Implikationen haben, die über herkömmliche Software-basierte Systeme hinausgehen. Als Gegenentwurf zur US-amerikanischen und chinesischen Konzeptualisierung von KI (‚Kommerz' beziehungsweise ‚Kontrolle') propagiert die Europäische Kommission die ‚vertrauenswürdige' KI. Diese soll zum einem Qualitätsmerkmal von KI ‚made in Europe' werden.

Das Konzept der vertrauenswürdigen KI ist dabei nicht neu, es orientiert sich an der Diskussion um ethische oder menschenzentrierte KI und nimmt Erkenntnisse aus dem Bereich der Mensch-Computer-Interaktion auf (Dragan 2019; Russell 2019; Shneiderman 2020). Das europäische Konzept wurde 2019 von der High-Level Expert Group on Artificial Intelligence der Europäischen Kommission entwickelt (HLEG 2019). Es umfasst die sieben Dimensionen:

- Vorrang menschlichen Handelns und menschlicher Kontrolle
- Technische Robustheit und Sicherheit
- Privatsphäre und Datenqualitätsmanagement
- Transparenz und Erklärbarkeit
- Vielfalt, Nichtdiskriminierung und Fairness
- Gesellschaftliches und ökologisches Wohlergehen
- Rechenschaftspflicht

Um diesen eher abstrakten Anforderungskatalog zu konkretisieren, hat die Expertengruppe 2020 eine Checkliste erarbeitet, mit der Anbieter und Nutzer überprüfen können, inwieweit das zu implementierende KI-System den Anforderungen der vertrauenswürdigen KI entspricht (HLEG 2020). ‚Vertrauenswürdige KI' bezeichnet dabei sowohl das Ziel als auch den Prozess, wie dieses Ziel erreicht werden kann.

In jüngster Zeit kamen weitere Vorschläge für die praktische Umsetzung und die ethische Bewertung von KI-Projekten hinzu und es wurden Arbeitsvorlagen, Checklisten und Guidelines vorgelegt. So hat zum Beispiel die AI Ethics Impact Group unter der Koordination des Verbands der Elektrotechnik Elektronik Informationstechnik e. V. und der Bertelsmann Stiftung ein KI-Ethik-Kennzeichnungssystem entwickelt, das sich an das Energieeffizienzlabel anlehnt und für die sechs Dimensionen Transparenz, Verantwortlichkeit, Privatheit, Gerechtigkeit, Zuverlässigkeit und ökologische Nachhaltigkeit entsprechende Güteklassen definiert (Hallersleben und Hustedt 2020).

Mitte 2021 befassten sich in Deutschland über 40 öffentlich geförderte Projekte mit dem Konzept vertrauenswürdiger KI.

Weitere Vorschläge zur konkreten Umsetzung wurden von der Plattform Lernende Systeme (Huchler et al. 2020) gemacht und in Forschungsprojekten wie *Ethik in der agilen Softwareentwicklung* des Bayerischen Forschungsinstituts für Digitale Transformation (Zuber et al. 2020). Weiterhin hat die Gesellschaft für Informatik (GI) ein Machine-Learning-Kennzeichnungssystem vorgelegt, das sich an der Idee des Beipackzettels von Medikamenten orientiert (Seifert et al. 2020). Für die Gestaltung des konkreten Projektablaufs und die Governancestruktur von KI-Entwicklungs- und Umsetzungsprojekten liegen ebenfalls Vorschläge vor (Shneiderman 2020; Puntschuh und Fetic 2020).

Doch welche Erfahrungen haben Unternehmen und Organisationen bisher mit den verschiedenen Ansätzen gemacht? Welche Guidelines haben sich als praxistauglich erwiesen, welche Herausforderungen gab es bei der Umsetzung? Da entsprechende Leitfäden und Governance-Empfehlungen seit einiger Zeit existieren, sollten sich erste konkrete Umsetzungsprojekte in Unternehmen und Organisationen finden lassen, die eine Auswertung von Erfahrungen im Hinblick auf Best Practices ermöglichen. Im Folgenden wird die Suche nach solchen Projekten nachgezeichnet und es werden die Ergebnisse vorgestellt.

Auf der Suche nach Beispielen für die praktische Umsetzung des Konzepts

Zunächst lässt sich feststellen, dass Beispiele für praktische Umsetzungen nicht auf der Hand liegen; es existieren keine Sammlungen oder Datenbanken und es gibt bisher auch keine einschlägigen Veröffentlichungen. Deshalb wurde eine Suchstrategie entwickelt, die im ersten Schritt Umsetzungsprojekte im akademischen Bereich (oft mit Beteiligung von Unternehmen) identifizierte. Im zweiten Schritt wurde Hinweisen auf Umsetzungen in großen Unternehmen nachgegangen und schließlich wurde die Start-up-Szene untersucht. Dazu wurden ausschließlich öffentlich zugängliche Quellen ausgewertet und keine exklusiven Zugänge in die Unternehmen hinein genutzt. Zwar wurden zusätzlich die einschlägigen Verbände gebeten, Umsetzungsbeispiele aus ihren Mitgliedsunternehmen zu benennen, jene haben von dieser Möglichkeit jedoch keinen Gebrauch gemacht.

Im akademischen Bereich, dem Startpunkt der Recherche, wurden die laufenden Forschungsprogramme des Bundes mit Bezug zu KI, die Aktivitäten ausgewählter Bundesländer sowie von Universitäten und Forschungsinstituten analysiert, um relevante Projekte zu identifizieren. Die Suchstrategie bestand darin, zunächst solche KI-Projekte zu identifizieren, in denen Aspekte der vertrauenswürdigen KI überhaupt eine Rolle spielen und anschließend auf Projekte zu fokussieren, die sich mit der konkreten Umsetzung von Richtlinien beschäftigen.

Ergebnis der Recherche ist, dass es Mitte 2021 in Deutschland über 40 öffentlich geförderte Projekte gibt, die sich mit der Umsetzung verschiedener Teilaspekte des Konzepts der vertrauenswürdigen KI, wie zum Beispiel der Transparenz und Erklärbarkeit befassen. In diesen Projekten wurden zum Teil auch Leitfäden entwickelt und adaptiert und in einigen Fällen in Kooperation mit Pilotunternehmen evaluiert. Allerdings stellte sich heraus, dass sich diese Projekte in einem frühen Stadium befinden und dass die bisher vorliegenden Ergebnisse noch keine systematische Zusammenschau von Strategien und Erfahrungen erlauben. Drei Projekte sollen diesen Befund illustrieren (Tabelle 1).

In einem zweiten Schritt wandte sich die Recherche dem Unternehmensbereich zu. Hier konnten mehrere große Firmen identifiziert werden, die sich mit ethischen Aspekten der von ihnen genutzten oder bereitgestellten KI-Systeme befassen. Tabelle 2 führt drei Beispiele auf. Dabei zeigt sich, dass die Beschäftigung dieser Unternehmen mit diesem Thema im Kontext ihrer jeweiligen Compliance- und Nachhaltigkeitsaktivitäten gesehen werden muss. Beispiele, wie konkrete Umsetzungen des Konzepts der vertrauenswürdigen KI in den Unternehmen vorgenommen wurden, konnten dagegen nicht recherchiert werden. Dies könnte neben den Gründen, die anschließend diskutiert werden, auch damit zusammenhängen, dass Unternehmen Details ihrer Softwareentwicklung nicht veröffentlichen, oder dass sie dies zum Schutz ihrer Kunden und Produkte bewusst vermeiden (Machmeier 2020).

Projekt	Ziele und Ergebnisse
ExamAI – KI Testing & Auditing	Entwicklung von Kontroll- und Prüfverfahren, mit denen KI-Systeme in industriellen Produktionssettings und im Personalmanagement sicher, transparent und diskriminierungsfrei gestaltet werden können sollen. Die Verfahren werden in Anwendungsszenarien getestet. Die Ergebnisse sollen es den Forschern dann ermöglichen, Empfehlungen für reale Implementierungen zu geben.
KI-Gestaltungsansätze und Ethik-Briefing der Plattform Lernende Systeme	Entwicklung umsetzungsorientierter Richtlinien und Change-Management-Empfehlungen für die Realisierung verantwortungsvoller KI-Systeme. Es werden einige Beispiele dafür angeführt, wie ausgewählte Aspekte in Unternehmen praktisch gehandhabt wurden. Die vorgeschlagenen Gestaltungsansätze sollen in Zukunft mit Leben gefüllt werden.
GOAL – Governance von und durch Algorithmen	Identifizierung von Governance-Strukturen und regulatorischem Handlungsbedarf für den verantwortungsvollen Einsatz von KI-Systemen. Ein Teilprojekt, das an der Westfälischen Wilhelms-Universität Münster durchgeführt wird, umfasst neben konzeptionellen Arbeiten, auch konkrete Tests und evaluiert die Erkenntnisse in einer Fallstudie mit einer selbst entwickelten mobilen App.

Tab. 1: Umsetzungsbeispiele aus dem Bereich der Forschung. *Quelle: eigene Darstellung*

Unternehmen	Aktivitäten mit Bezug auf vertrauenswürdige KI
Deutsche Telekom	Entwicklung eines internen AI-Code-of-Conduct. Für die Umsetzung wurde eine Prüfmatrix zur ethischen Bewertung neuer KI-Produkte und ein internes Gütesiegel entwickelt. Das Unternehmen hat angekündigt, mit Wirtschaftsprüfern an einer Zertifizierung zu arbeiten, mit der der ethische und vertrauenswürdige Einsatz von KI im Unternehmenskontext nachgewiesen werden soll (Mackert 2020).
SAP	Veröffentlichung eines Handbuchs zur KI-Ethik, das auf der Arbeit eines internen KI-Ethik-Lenkungsausschusses und eines externen KI-Ethik-Beirats basiert (Heesen et al. 2020, S. 26). Hinsichtlich konkreter Erfahrungen bei Implementierungsprojekten ist das Unternehmen zurückhaltend: „Aus Rücksicht auf die beteiligten Kunden und Kollegen ist es schwierig, über konkrete Szenarien zu sprechen" (Markus Noga, zitiert in Machmeier 2020).
BMW	Ankündigung einer generellen Verankerung der sieben EU-Anforderungen für vertrauenswürdige KI im Unternehmen. Im Hinblick auf konkrete Ansätze oder Erfahrungen mit der Umsetzung des Konzepts konnten keine Berichte gefunden werden.

Tab. 2: Aktivitäten großer Unternehmen mit Bezug zur Umsetzung vertrauenswürdiger KI. *Quelle: eigene Darstellung*

Start-ups	Aktivitäten mit Bezug auf vertrauenswürdige KI
Datanizing	Die Algorithmen der Firma Datanizing segmentieren Konsumenten auf Basis von Daten aus sozialen Netzwerken und stellen das Ergebnis Marktforschungsunternehmen zur Verfügung. Die Gründer wollen ihren Kunden ein besseres Verständnis für die Potenziale und Grenzen der verwendeten Algorithmen ermöglichen und haben einen Leitfaden entwickelt, der dabei helfen soll, die soziale Dimension zu berücksichtigen (Heesen et al. 2020. S. 25).
Prenode	Das Start-up Prenode hat eine Software entwickelt, die es ermöglicht, Daten aus verschiedenen Quellen zu verarbeiten, ohne sie in einer zentralen Datenbank zusammenführen zu müssen. Die Zusammenführung ist aufgrund von Datenschutzbedenken oft problematisch (techtag 2020). Die gefundene Lösung ist insbesondere für den Gesundheitsbereich interessant, in dem Patientendaten aus verschiedenen Datenbanken verarbeitet werden müssen.

Tab. 3: Umsetzungsbeispiele von Start-ups. *Quelle: eigene Darstellung*

Im dritten Rechercheschritt wurde nach Umsetzungsbeispielen bei Start-ups und mittelständischen Unternehmen, insbesondere in der IT- und Softwarebranche gesucht. Doch dieser Rechercheschritt erwies sich als am wenigsten produktiv. Start-ups und mittelständische Unternehmen scheinen derzeit abzuwarten, wie sich das Thema entwickelt. Dennoch sollen hier zwei Beispiele angeführt werden, die zeigen, dass bestimmte Aspekte des Themas auch in der Start-up-Szene eine Rolle spielen (Tabelle 3).

Insgesamt zeigt die Recherche, dass es trotz einiger Ausnahmen eine große Lücke zwischen den konzeptionellen Angeboten und der praktischen Umsetzung gibt. Das Umsetzungsdefizit ist dabei nicht auf Deutschland beschränkt. Im The AI Index 2021 Annual Report der Stanford University schreiben die

Autoren, dass sie überrascht waren, wie wenig sie zum Thema Umsetzung vertrauenswürdiger KI gefunden haben: „Though a number of groups are producing a range of qualitative or normative outputs in the AI ethics domain, the field generally lacks benchmarks“ (Zhang et al. 2021, S. 127). Es ist daher naheliegend zu fragen, wodurch diese Lücke, beziehungsweise der Mangel an Umsetzungsprojekten zustande kommt. Im folgenden Abschnitt werden drei mögliche Erklärungen aufgeführt.

Gründe für das Umsetzungsdefizit

Die folgenden Erklärungen basieren neben der Analyse einschlägiger Literatur auf der Auswertung von Interviews, die im April und Mai 2021 mit zehn Expert:innen durchgeführt wurden.

Der erste Grund für das Umsetzungsdefizit hat mit Time-to-Market-Überlegungen der Unternehmen zu tun: *First Mover* (Pionierunternehmen) scheinen auf dem noch relativ jungen KI-Markt einen erheblichen Marktvorteil zu haben, auch wenn ihre KI-Anwendung bis zu einem gewissen Grad fehlerbehaftet ist. Ein Beispiel hierfür ist die Personalauswahl-KI des Münchner Startups Retorio. Obwohl die Anwendung nachweislich bestimmte ethnische Gruppen diskriminiert und anfällig für Manipulationen ist, scheint sie viele Kunden gefunden zu haben, da sie eine der ersten Anwendungen auf dem Markt war (Radü 2021).

Ein zweiter Grund ist, dass es für Entwickler und Manager oft schwierig ist zu entscheiden, inwiefern das Konzept der vertrauenswürdigen KI für ihre spezifische Anwendung überhaupt relevant ist. Ein großer Teil der aktuellen KI-Anwendungen – so scheint es – wird in einem Fertigungskontext (vorausschauende Wartung, Maschinen- und Logistikoptimierung, neue Materialien) oder in einem Forschungskontext (Modelloptimierung, Visualisierung) eingesetzt. Robustheit und Sicherheit sind hier wichtige Aspekte. Vielfalt, Nichtdiskriminierung und Fairness spielen aber nur zum Teil eine Rolle. Auf der anderen Seite gibt es KI-Systeme, die als entscheidungsunterstützende Systeme bei Kredit- oder Versicherungsunternehmen eingesetzt werden, im Personalmanagement, in öffentlichen Verwaltungen (Sozialamt, Arbeitsamt) und im Gesundheitswesen. Hier spielen ethische Fragen und Persönlichkeitsrechte des Einzelnen eine sehr wichtige Rolle. Tatsächlich ist das Anwendungsspektrum von KI generell sehr breit. KI sollte daher nicht nur im Kontext automatisierter Entscheidungssysteme gesehen werden (Köszegi 2020), sondern auch im Kontext von Muster- und Bilderkennung, von Prozessoptimierungen sowie von Empfehlungssystemen oder auch dem autonomen Fahren. Für all diese Anwendungen gilt es derzeit, jeweils spezifisch zu entscheiden, welche Aspekte der vertrauenswürdigen KI prioritär berücksichtigt werden müssen und welche eine weniger wichtige Rolle spielen.

Der dritte Grund bezieht sich auf die Tatsache, dass KI-Projekte nicht wie ‚normale‘ Software-Projekte umgesetzt werden können. Die Fähigkeit von KI-Systemen, ihre Outputs kontinuierlich an veränderte Inputs anzupassen (‚Lernen‘), macht gerade ihren Reiz aus (Heesen 2021). Diese Anpassungsfähigkeit unterscheidet KI von anderen Softwareprogrammen, die zwar zum Teil auch an unterschiedliche Kontexte angepasst werden können, deren Verarbeitungsgrundlagen sich aber nicht grundsätzlich durch neu eingespeiste Daten verändern. Im Vergleich zur Standard-Softwareentwicklung erfordert die Entwicklung und Implementierung vertrauenswürdiger KI-Anwendungen zusätzliche Vorkehrungen: Auf der Ebene der Softwareentwicklung beinhaltet dies die Einführung spezifischer Prüfungsprozesse und Bias-Tests sowie die Integration von Erklärungen und auf der Management-Ebene die Einführung von internen Überprüfungen, kontinuierlichen Fehleranalysen sowie speziellen Strukturen der Aufsicht und Kontrolle (Shneiderman 2020, S. 12). Derartige, neue Arbeitsabläufe und Prüfprozesse zu implementieren ist eine Herausforderung für Firmen und Organisationen, denn sie erfordern entsprechendes Know-how, spezifische Planung und zusätzliche Ressourcen.

Fazit und Ausblick: Wege aus dem Umsetzungsdefizit

Welche Möglichkeiten gibt es nun, das Umsetzungsdefizit perspektivisch zu verringern und mehr Unternehmen und Organisationen dazu zu motivieren, das Konzept der vertrauenswürdigen KI auch anzuwenden? Basierend auf den Erkenntnissen des vorangegangenen Abschnitts werden im Folgenden zwei mögliche Ansätze vorgeschlagen. Der erste bezieht sich auf die weitere Konkretisierung von Leitfäden und die Einführung von Zertifikaten, der zweite auf die Einrichtung gemischter Teams als Teil einer spezifischen KI-Governance-Struktur.

First Mover scheinen auf dem noch relativ jungen KI-Markt einen erheblichen Marktvorteil zu haben.

Konkretisierung von Leitfäden und Einführung von Zertifikaten

Obwohl es heute eine Vielzahl von Leitfäden und konkreten Umsetzungsvorschlägen gibt, reicht der Grad der Konkretisierung für die komplexe Implementation von KI-Projekten vielfach nicht aus. Was fehlt, sind Überlegungen zur Art und Weise, wie Daten generiert und verarbeitet werden, wie Trainingsdatensätze ausgewählt und wie geeignete Algorithmen bestimmt werden können. Hagendorff (2020, S. 111) verweist in diesem Zusammenhang darauf, dass es bei der Konkretisierung von Leitfä-

den darauf ankommt, Ethik in ‚Mikroethik' zu verwandeln, d. h. die ethische Debatte auf ein handhabbares Konkretisierungsniveau herunterzubrechen.

Ein weiterer Ansatz, um die Verbreitung vertrauenswürdiger KI zu erhöhen, ist die Einführung von Zertifikaten. Die prominentesten Zertifizierungsaktivitäten in Deutschland sind derzeit die Normungsroadmap zur künstlichen Intelligenz, erarbeitet durch den DIN e. V. und das Bundesministerium für Wirtschaft und Energie sowie die Plattform zur Zertifizierung von KI-Anwendungen des Landes Nordrhein-Westfalen. Darüber hinaus hat der KI-Bundesverband ein KI-Gütesiegel herausgegeben, zu dem sich die Mitglieder des Verbandes in einer freiwilligen Vereinbarung verpflichten können. In allen drei Initiativen sollen künftig konkrete Anwendungsfälle in Zusammenarbeit mit Unternehmen entstehen (Schonschek 2020). Mit entsprechenden Zertifikaten können Unternehmen für ihre Anwendungen werben. Außerdem erhöhen Zertifikate das Bewusstsein in der Öffentlichkeit für die Notwendigkeit, sich mit Aspekten der vertrauenswürdigen KI zu beschäftigen.

Zertifikate erhöhen das Bewusstsein für die Notwendigkeit, sich mit Aspekten der vertrauenswürdigen KI zu beschäftigen.

Einrichtung gemischter Teams

Wie erwähnt, erfordert die Umsetzung des Konzepts der vertrauenswürdigen KI spezifische Abläufe und zusätzliche Prüfprozesse. Dabei kommt es darauf an, die beiden Welten der Softwareentwicklung und der Ethik zusammenzubringen. Hierfür eignen sich insbesondere gemischte Teams, denn Software-Ingenieure und Informatiker:innen überblicken in der Regel nicht die Literatur und die Empfehlungen ethischer Diskussionen, und Ethiker:innen und Sozialwissenschaftler:innen sind in der Regel nicht mit den Anforderungen der Programmierung oder organisatorischen Details vertraut. Der Ruf nach interdisziplinären Teams ist dabei nicht neu (Shneiderman 2020; Hagendorff 2020; Lopez 2021; Kusner und Loftus 2020). Allerdings gibt es hier mit Ausnahme einiger Projekte in der Gesundheitsforschung (Franke 2021) noch zu wenige Umsetzungsbeispiele.

Projekte mit gemischten Teams könnten die Umsetzungsbilanz verbessern und die Diskussion um Best Practices und Erfolgsstrategien bereichern. Entsprechende Methoden aus dem Bereich der Science Technology Studies, wie zum Beispiel die der *Embedded Social Scientists* liegen vor (Fisher und Schuurbiers 2013; Simone 2018; Gransche und Manzeschke 2020; Haux und Karafyllis 2021) und könnten auf den Bereich der vertrauenswürdigen KI übertragen werden.

Best-Practice-Darstellungen und Erfolgsbeispiele sind dringend notwendig, um die Vorteile vertrauenswürdiger KI-Systeme zu demonstrieren und Nachahmer zu motivieren. Unternehmen und Organisationen sollten ein Interesse daran haben, ihre Bemühungen um die Vertrauenswürdigkeit ihrer Anwendungen auch zu kommunizieren, da dies die Attraktivität und Akzeptanz ihrer Angebote prinzipiell erhöht.

Angabe von Finanzierungsquellen

Dieser Artikel wurde aus Mitteln der KI-Gruppe am Fraunhofer ISI (www.isi.fraunhofer.de/de/themen/ki.html) finanziert.

Literatur

Dragan, Anca (2019): Der unberechenbare Mensch. In: Süddeutsche Zeitung, 23. 04. 2019. Online verfügbar unter https://www.sueddeutsche.de/kultur/kuenstliche-intelligenz-roboter-realitaet-1.4418243?reduced=true, zuletzt geprüft am 08. 10. 2021.

Fisher, Erik; Schuurbiers, Daan (2013): Socio-technical integration research. Collaborative inquiry at the midstream of research and development. In: Neelke Doorn et al. (Hg.): Early engagement and new technologies. Opening up the laboratory. Dordrecht: Springer, S. 97–110. https://doi.org/10.1007/978-94-007-7844-3_5

Franke, Thomas (2021): Kooperative und kommunizierende KI-Methoden für die medizinische bildgeführte Diagnostik. Webdarstellung des Projekts CoCoAI. Online verfügbar unter https://www.imis.uni-luebeck.de/de/forschung/projekte/cocoai, zuletzt geprüft am 08. 10. 2021.

Gransche, Bruno; Manzeschke, Arne (Hg.) (2020): Das geteilte Ganze. Horizonte integrierter Forschung für künftige Mensch-Technik-Verhältnisse. Wiesbaden: Springer. https://doi.org/10.1007/978-3-658-26342-3

Hagendorff, Thilo (2020): The ethics of AI ethics. An evaluation of guidelines. In: Minds & Machines 30, S. 99–120. https://doi.org/10.1007/s11023-020-09517-8

Hallersleben, Sebastian; Hustedt, Carla (2020): From principle to practice. An interdisciplinary framework to operationalise AI ethics. Frankfurt a. M.: VDE & Bertelsmann Stiftung. Online verfügbar unter https://www.ai-ethics-impact.org/resource/blob/1961130/c6db9894ee73aefa489d6249f5ee2b9f/aieig---report---download-hb-data.pdf, zuletzt geprüft am 08. 10. 2021.

Haux, Reinhold; Karafyllis, Nicole (2021): Methodisch-technische Aspekte der Evaluation erweiterten Zusammenwirkens. In: Reinhold Haux, Klaus Gahl, Meike Jipp und Otto Richter (Hg.): Zusammenwirken von natürlicher und künstlicher Intelligenz. Wiesbaden: Springer VS Open Access, S. 175–198. https://doi.org/10.1007/978-3-658-30882-7_13

Heesen, Jessica (2021): Wie kommt Ethik in die Künstliche Intelligenz? In: Digitale Welt, 06. 01. 2021. Online verfügbar unter https://digitaleweltmagazin.de/2021/01/06/wie-kommt-ethik-in-die-kuenstliche-intelligenz/, zuletzt geprüft am 19. 10. 2021.

Heesen, Jessica; Grunwald, Armin; Matzner, Tobias; Roßnagel, Alexander (2020): Ethik-Briefing. Leitfaden für eine verantwortungsvolle Entwicklung und Anwendung von KI-Systemen. München: Lernende Systeme – Die Plattform

für Künstliche Intelligenz. Online verfügbar unter https://www.plattform-lernende-systeme.de/files/Downloads/Publikationen/AG3_Whitepaper_EB_200831.pdf, zuletzt geprüft am 08. 10. 2021.

HLEG – High-Level Expert Group on Artificial Intelligence (2019): Ethics guidelines for trustworthy AI. Brussels: European Commission. Online verfügbar unter https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60419, zuletzt geprüft am 19. 10. 2021.

HLEG (2020): The assessment list for Trustworthy Intelligence (ALTAI) for self-assessment. Brussels: European Commission. Online verfügbar unter https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment, zuletzt geprüft am 08. 10. 2021.

Huchler, Norbert et al. (2020): Kriterien für die menschengerechte Gestaltung der Mensch-Maschine-Interaktion bei Lernenden Systemen. Online verfügbar unter https://www.plattform-lernende-systeme.de/files/Downloads/Publikationen/AG2_Whitepaper2_220620.pdf, zuletzt geprüft am 08. 10. 2021.

Köszegi, Sabine (2020): Der autonome Mensch im Zeitalter des Digitalen Wandels. In: Markus Hengstschläger (Hg.): Digitaler Wandel und Ethik. München: Ecowin Verlag, S. 62–90.

Kusner, Matt; Loftus, Joshua (2020): The long road to fairer algorithms. Build models that identify and mitigate the causes of discrimination. In: Nature 578, S. 34–36. https://doi.org/10.1038/d41586-020-00274-3

Lopez, Paola (2021): Artificial Intelligence und die normative Kraft des Faktischen. In: Merkur 75 (863), S. 42–52.

Machmeier, Corinna (2020): Verantwortungsvoll mit Künstlicher Intelligenz umgehen. Ein Jahr des Lernens. In: SAP News Center, 09. 01. 2020. Online verfügbar unter https://news.sap.com/germany/2020/01/ki-kuenstliche-intelligenz-ethik/, zuletzt geprüft am 08. 10. 2021.

Mackert, Manuela (2020): Compliance und künstliche Intelligenz. Im Blickpunkt – Warum sich Unternehmen einen Code-of-Conduct leisten sollten. In: Compliance Business 1, S. 3–5. Online verfügbar unter https://www.deutscheranwaltspiegel.de/compliancebusiness/kuenstliche-intelligenz/compliance-und-kuenstliche-intelligenz-19597/, zuletzt geprüft am 20. 10. 2021.

Puntschuh, Michael und Fetic, Lajla (2020): Praxisleitfaden zu den Algo.Rules. Orientierungshilfen für Entwickler:innen und ihre Führungskräfte. Gütersloh: Bertelsmann Stiftung. Online verfügbar unter https://www.bertelsmann-stiftung.de/fileadmin/files/alg/Algo.Rules_Praxisleitfaden.pdf, zuletzt geprüft am 19. 10. 2021.

Radü, Jens (2021): Wie Künstliche Intelligenz über Ihren nächsten Job entscheidet. In: Der Spiegel, 08. 03. 2021, S. 68–70.

Russell, Stuart (2019): Human compatible. Artificial Intelligence and the problem of control. New York: Viking.

Schonschek, Oliver (2020): Künstliche Intelligenz muss Vertrauen schaffen. In: Com! professional, 20. 01. 2020. Online verfügbar unter https://www.com-magazin.de/praxis/kuenstliche-intelligenz/kuenstliche-intelligenz-vertrauen-schaffen-2451438.html, zuletzt geprüft am 19. 10. 2021.

Seifert, Christin; Scherzinger, Stefanie; Wiese Lena (2020): Beipackzettel für Modelle des maschinellen Lernens für fairere KI. In: Gesellschaft für Informatik e. V. (Hg.): Themen in der GI, 27. 11. 2020. Online unter https://gi.de/themen/beitrag/beipackzettel-fuer-modelle-des-maschinellen-lernens-fuer-fairere-ki, zuletzt geprüft am 19. 10. 2021.

Shneiderman, Ben (2020): Bridging the gap between ethics and practice. Guidelines for reliable, safe, and trustworthy human-centered AI Systems. In: ACM Transactions on Interactive Intelligent Systems 10 (4), S. 1–31. https://doi.org/10.1145/3419764

Simone, Angela (2018): Steering research and innovation through RRI. What horizon for Europe? In: Journal of Science Communication 3, 6S. https://doi.org/10.22323/2.17030302

techtag Redaktion (2020): prenode im Gründerview. 10 Fragen an Dr.-Ing. Robin Hirt und Dr. Ronny Schüritz. Online verfügbar unter https://www.techtag.de/startups/gruenderview/prenode-im-gruenderview/, zuletzt geprüft am 08. 10. 2021.

Zhang, Daniel et al. (2021): The AI Index 2021 Annual Report. Stanford, CA: Human-Centered AI Institute, Stanford University. Online verfügbar unter https://aiindex.stanford.edu/wp-content/uploads/2021/03/2021-AI-Index-Report_Master.pdf, zuletzt geprüft am 20. 10. 2021.

Zuber, Niina; Kacianka, Severin; Pretschner, Alexander; Nida-Rümelin, Julian (2020): Ethische Deliberation für agile Softwareprozesse. EDAP-Schema. In: Markus Hengstschläger (Hg.): Digitaler Wandel und Ethik. München: Ecowin Verlag, S. 160–184.



DR. BERND BECKERT

ist stellvertretender Leiter des Competence Centers (CC) Neue Technologien am Fraunhofer-Institut für System- und Innovationsforschung (ISI). Er ist zudem Leiter der KI-Gruppe am ISI, die im Jahr 2020 gegründet wurde. Die KI-Gruppe beschäftigt sich CC-übergreifend mit künstlicher Intelligenz aus einer TA- und Innovationsforschungs-Perspektive.