10

SPECIAL TOPIC

KI-Systeme gestalten und erfahren

Konzepte, Werte, Anwendungen

Künstliche Intelligenz (KI) ist derzeit in Wirtschaft, Gesellschaft und Wissenschaft präsent. Aufgabe der Technikfolgenabschätzung ist es, öffentliche Erwartungen und Befürchtungen in sachliche, realistische Perspektiven zu transformieren sowie Impulse für eine wünschenswerte Gestaltung von KI zu geben.

Herausgegeben von Bernhard G. Humm, Stephan Lingner, Jan C. Schmidt und Karsten Wendland

https://doi.org/10.14512/tatup.30.3.10

EINLEITUNG

KI-Systeme

Aktuelle Trends und Entwicklungen aus Perspektive der Technikfolgenabschätzung

Bernhard G. Humm, *Hochschule Darmstadt – University of Applied Sciences, Fachbereich Informatik, Haardtring 100, 64295 Darmstadt, DE (bernhard.humm@h-da.de)* *0000-0001-7805-1981*
Stephan Lingner, *Institut für qualifizierende Innovationsforschung und -beratung GmbH (IQIB), Bad Neuenahr-Ahrweiler, DE (stephan.lingner@iqib.de)*
Jan C. Schmidt, *Hochschule Darmstadt – University of Applied Sciences, Hochschule Darmstadt, Fachbereich Gesellschaftswissenschaften, Darmstadt, DE (jan.schmidt@h-da.de)*
Karsten Wendland, *Karlsruher Institut für Technologie (KIT), Institut für Technikfolgenabschätzung und Systemanalyse (ITAS), Karlsruhe, DE (karsten.wendland@kit.edu)* *0000-0003-4812-0928*

Zusammenfassung • Künstliche Intelligenz (KI) ist – ebenso wie damit verknüpfte Techniken wie maschinelles Lernen und Big Data – in aller Munde. Die große Dynamik und Tragweite dieser Entwicklungen zeigen sich bereits in zahlreichen Anwendungsgebieten von Wirtschaft, Gesellschaft und Wissenschaft. Technikfolgenabschätzung (TA) von KI hat in diesem Zusammenhang zunächst die Aufgabe, etwaige überzogene öffentliche Erwartungen und Befürchtungen in sachliche, realistische Perspektiven zu transformieren. In einem zweiten Schritt kann TA entlang begründbarer Entwicklungsziele von KI und legitimer gesellschaftlicher Wertvorstellungen Impulse für die weitere, wünschbare Gestaltung von KI geben. Wenn TA diese Orientierungsaufgabe nah am technologischen Kern wahrnimmt, findet sie dabei große gestalterische Freiräume in frühen Phasen der Technikentwicklung vor. Die damit zusammenhängenden Gedanken werden im vorliegenden Einleitungskapitel konkretisiert und auf die Beiträge zu diesem Themenschwerpunkt angewendet.

Designing and experiencing AI systems. *Recent trends and developments from a technology assessment perspective*

Abstract • *Artificial intelligence (AI) is on everyone's lips – as well as the associated technologies of machine learning and big data. The enormous dynamics and consequences of these developments become already evident in numerous areas of application in business, society and science. In this context, technology assessment (TA) of AI initially has the task of transforming any excessive public expectations and fears to the factual level. In a second step, TA can provide impulses for the further, desirable design of AI based on reasonable development goals of AI and legitimate societal values. If TA conducts this orientation task close to the technological core, it can consider wide scopes of options for action in the early phases of technology development. Related thoughts are put into concrete terms in this article and will be related to the authors' contributions to this topical focus.*

Keywords • *Artificial Intelligence, AI, Machine Learning, Big Data, Technology Assessment*

Künstliche Intelligenz (KI) und die mit ihr verbundenen Ermöglichungstechniken wie maschinelles Lernen und Big Data sind derzeit von einer großen öffentlichen Aufmerksamkeit gekennzeichnet. Die hohe Entwicklungsdynamik ist zudem mit einem erheblichen Veränderungspotenzial für Wirtschaft, Gesellschaft und Wissenschaft verknüpft. In der öffentlichen Debatte kursieren zudem teilweise unreflektierte Erwartungen an KI und auch überzogene Befürchtungen vor ihrem breiten Einsatz. Angesichts ihrer möglichen Konsequenzen für alle gesellschaftlichen Lebensbereiche erscheint es daher geboten, Erforschung, Entwicklung und Einsatz von KI entlang begründbarer Zielvorstellungen und legitimer Wertvorstellungen zu rahmen und zu gestalten. Dies ist eine Aufgabe der Technikfolgenabschätzung (TA), die hier zur Klärung und Orientierung auf unsicherem und ambivalentem Terrain in ähnlicher Weise gefragt ist, wie bei anderen tiefgreifenden Entwicklungen in Bereichen der Nanotechnologie oder der Biotechnologie. Technikfolgenabschätzung von KI betrifft dabei nicht nur die Implementierung und Anwendungsfelder von bereits entwickelten KI-Technologien, sondern insbesondere auch die vorgelagerte, im öffentlichen Diskurs oft vernachlässigte Ebene der Forschung. In diesen Frühphasen der

https://doi.org/10.14512/tatup.30.3.11
Received: Oct. 27, 2021; revised version accepted: Nov. 02, 2021; published online: Dec. 20, 2021 (editorial peer review)

Technikentwicklung ist TA allgemein mit großen gestalterischen Freiräumen nah am technikwissenschaftlichen Kern von KI konfrontiert, aber eben auch mit besonderen methodischen Herausforderungen der Vorausschau ihrer Folgen (Collinridge 1982).

So ist hier zunächst die Suche nach geeigneten Ansätzen *für eine prospektive Beurteilung von KI* zu nennen. Dabei stellt sich unter anderem die Frage, wie retrospektive Analysen von bereits abgeschlossenen – und somit evidenten – Digitalisierungsdebatten für heutige Ideen für KI-Zukünfte nutzbar gemacht werden können.

Für die gestaltungsorientierte Beurteilung von KI ist sicher von zentraler Bedeutung, wie sich die *Autonomie und Kreativität sowie die Komplexität und Opazität* künstlicher ‚intelligenter' Systeme im gesellschaftlichen Alltag und in der Forschungspraxis aus TA-Sicht darstellt. Hier sind zum Beispiel spezifische Probleme künstlicher Kommunikation durch technische Agenten in sozialen Medien zu nennen. Ein weiterer Punkt ist die Frage, wie vertrauenswürdige KI entwickelt und in die Praxis umgesetzt werden kann und welche Hemmnisse dabei zu überwinden sind.

Daran schließt sich unmittelbar auch die Notwendigkeit der Gestaltung von geeigneten Rahmenbedingungen für den Einsatz von KI-Systemen in der gesellschaftlichen Praxis an. Hier sind zunächst *aussichtsreiche Regulierungsansätze*, ihre ‚offenen Flanken' sowie mögliche Lösungsvorschläge zu erarbeiten, die wünschbare KI-Innovationen fördern können. Diese Vorschläge sollten dabei auch auf die unterschiedlichen immanenten Geschwindigkeiten gesetzgeberischer und technischer Entwicklungen eingehen.

Ein anderes praktisches Problem ist die Frage, wie *notwendige KI-Kompetenzen* für den aufgeklärten Einsatz entsprechender Systeme in der Bevölkerung vermittelt und verankert werden können. Dieses Anliegen richtet sich darauf, in der Breite der Gesellschaft ein KI-Verständnis zu stärken, das technischen Laien erlaubt, Möglichkeiten und Risiken von KI-Systemen besser einschätzen können, um ihre Handlungs- und Entscheidungskompetenzen auch zukünftig einsetzen zu können. Auf Unternehmensseite stellt sich auch die Frage, wie der mitarbeiterfreundliche Einstieg in betriebliche KI-Anwendungen gelingen kann und welche flankierenden Forschungsansätze hierzu beitragen können. Diese sind insbesondere auf die Analyse und Abschätzung kritischer *Vertrauens- und Akzeptanzbedingungen* hin zu entwickeln.

Die vorstehenden Überlegungen sind aber auch auf konkrete Problemkontexte hin zu beziehen und für mögliche Lösungsansätze zu ertüchtigen; mögliche Anwendungsfälle aus vielen Beispielen sind die Abschätzung der Chancen und Risiken von KI in der biomedizinischen Diagnose und Therapie von Krankheiten.

Die folgenden Abschnitte erläutern den Kern von KI-Systemen und ihrer Nutzbarkeit vor dem Hintergrund ihrer öffentlichen Wahrnehmung. Es folgt sodann eine kursorische Übersicht der Themenbeiträge dieses Schwerpunkts mit Vorschlägen für die Gestaltung von KI-Systemen und ihres Anwendungsrahmens.

KI – Hype und Realität[1]

Die hohe Öffentlichkeitswirksamkeit von KI ist mit überzogenen Erwartungen einerseits sowie mit übertriebenen Befürchtungen andererseits verknüpft. Hinzu kommt, dass KI ein beliebtes Thema für Kinofilme ist und so die Gefahr besteht, dass Fiktion und Wirklichkeit im Bewusstsein vieler verschwimmen.

Herbert Simon, einer der Gründerväter von KI meinte bereits 1965: „Machines will be capable […] of doing any work that a man can do" (Allen 2001). Und sein Kollege Marvin Minski prophezeite 1970: „Within 10 years computers won't even keep us as pets" (Allen 2001). Keine dieser Prophezeiungen hat sich auch nur annähernd erfüllt. Doch das hält zeitgenössische Autor*innen und Visionär*innen auch heute nicht davon ab, überzogenen Erwartungen und Befürchtungen medienwirksam auszurufen. Prominente Vertreter gewagter Positionen sind Ray Kurzweil, Nick Bostrom, Yuval Noah Harai, Elon Musk, sowie der verstorbene Stephen Hawking. So behauptet Ray Kurzweil (2015, S. 9): „Mit der Singularität werden wir die Grenzen unserer biologischen Körper und Gehirne überschreiten. Wir werden die Gewalt über unser Schicksal erlangen. Unsere Sterblichkeit wird in unseren Händen liegen. Wir werden so langen leben können wie wir wollen […]. Bis zum Ende des Jahrhunderts wird die nichtbiologische Komponente unserer Intelligenz Trillionen Mal mächtiger sein als bloße menschliche Intelligenz".

Die hohe Öffentlichkeitswirksamkeit von KI ist mit überzogenen Erwartungen einerseits sowie mit übertriebenen Befürchtungen andererseits verknüpft.

Die meisten professionellen Einschätzungen von KI sind demgegenüber deutlich nüchterner. Uns Gastherausgebern ist Stand heute keinerlei Evidenz bekannt, dass sogenannte ‚starke KI', also KI-Anwendungen, die sich, wie oben prognostiziert,

1 Dieser Abschnitt ist in weiten Teilen sinngemäß der Studie von Gethmann et al. (2021) entnommen, an der Bernhard Humm, Stephan Lingner und Jan Schmidt maßgeblich mitgewirkt haben.

eigenständig weiterentwickeln, wirklich möglich sind. In jedem Fall sind diese aus unserer Sicht nicht in greifbarer Nähe. Besser werden lediglich die Imitationen und Inszenierungen von Intelligenz. Auch der hohe Anspruch, dereinst KI mit Bewusstsein zu erschaffen, geht stark mit der Reduzierung des Bewusstseinsbegriffs auf Funktionalitäten einher, nicht aber mit etwaigen Erlebnisqualitäten. Alle KI-Anwendungen, die bis heute entwickelt wurden, werden der sogenannten ‚schwachen KI' zugeordnet. Sie sind auf spezielle Aufgaben zugeschnitten, deren Lösung bislang menschliche Fähigkeiten erforderten, und lösen diese selbsttätig.

Klar ist, dass KI bereits heute eine hohe Bedeutung in Wirtschaft und Gesellschaft hat und diese Bedeutung voraussichtlich weiter zunehmen wird. Für viele Unternehmen hat KI eine strategische Bedeutung. So stammt beispielsweise folgende Aussage von Amazon: „Without ML [machine learning], Amazon.com couldn't grow its business, improve its customer experience and selection, and optimize its logistic speed and quality." (Marr 2018). Ähnlich äußern sich Google, Facebook, IBM und andere Technologie-Konzerne (Marr 2018). Auch die deutsche Bundesregierung hat 2018 eine KI-Strategie (BMBF 2021) verabschiedet mit dem Ziel, die Erforschung, Entwicklung und Anwendung von künstlicher Intelligenz in Deutschland auf ein weltweit führendes Niveau zu bringen und zu halten.

Perspektiven von KI

KI steht in der Tradition von Automatisierungs-Technologien, bei der menschliche Arbeit von Maschinen übernommen wird. Besonders seit der industriellen Revolution verändert diese Automatisierung in rasanter Weise alle Lebens- und Arbeitsbereiche. Viele Berufsgruppen sind komplett verschwunden, neue sind entstanden, und Menschen sowie Gesellschaften mussten sich stets an veränderte Gegebenheiten anpassen, die durch technische Entwicklungen verursacht wurden. Während im Zuge der industriellen Revolution vorwiegend physische Tätigkeiten von Maschinen übernommen wurden, werden im Zuge der Digitalisierung, insbesondere mit KI-Techniken, zunehmend auch kognitive Tätigkeiten automatisiert. Damit verbunden sind Chancen und Risiken, die oft eng beieinander liegen, sich sogar gegenseitig bedingen können. Wir möchten dies anhand von Anwendungsbeispielen erläutern.

KI wird zunehmend in der *Medizin* eingesetzt; so werden beispielsweise bildgebende Verfahren mittels Röntgenstrahlung, Ultraschall und MRT in Kombination mit KI-Verfahren zur (semi-)automatischen Analyse beziehungsweise Diagnose eingesetzt. Diese übertreffen in speziellen Gebieten teilweise sogar menschliche Ärzt*innen. Die Chancen hier liegen in der verbesserten Diagnostik, einer höheren Reichweite (z. B. bis hin zu Hausarztpraxen) und reduzierten Kosten im Gesundheitswesen. Demgegenüber steht das Risiko von Fehldiagnosen z. B. aufgrund von Verzerrungen (bias) in den Trainingsdaten für machine learning-Anwendungen. Ironischerweise kann dies eine indirekte Folge der Chance verbesserter Diagnostik sein, wenn über einen längeren Zeitraum der erfolgreichen Verwendung einer KI-Anwendung sich schleichend ein blindes Vertrauen einstellt, insbesondere bei einer neuen Generation von Ärzt*innen. Dann kann schrittweise die Expertise zur menschlichen Diagnose, welche der maschinellen Diagnose kritisch gegenübergestellt wird, sinken.

Als nächstes Beispiel betrachten wir das sogenannte *autonome Fahren*, derzeit noch in der Experimentierphase, aber mit Ausblick auf breiten Einsatz in vielen Ländern. Die Chancen bestehen in einer höheren Effizienz und damit verbunden geringeren Kosten, z. B. durch Wegfall von Taxi-Fahrer*innen. Man kann auch hoffen, dass mit der Automatisierung car sharing-Konzepte umgesetzt werden können, sodass sich perspektivisch die Anzahl der Fahrzeuge, somit auch die damit verbundene Umweltbelastung sowie die Verkehrsverdichtung der Städte reduziert. Mit fortschreitender Technik könnten autonome Fahrzeuge im Durchschnitt sicherer fahren als menschliche Fahrer*innen. Allerdings bleibt auch bei immer besserer Technik grundsätzlich ein Risiko von Unfällen, verursacht durch ‚autonome' Fahrzeuge. Hier bestehen noch große ungeklärte juristische Fragen, z. B. nach der Verantwortung und Haftung. Die Sicherheit kann aber auch drastisch sinken, da mit durchgängigem autonomem Fahren das Risiko großflächiger, verheerender Cyberangriffe im Verkehr steigt. Auch die Hoffnung auf eine Reduzierung der Umweltbelastung könnte sich in das Gegenteil umkehren, wenn die Möglichkeit von einfach verfügbarem Individualtransport mit ‚autonomen' Fahrzeugen den Bedarf nach Mobilität ständig wachsen lässt – ein sogenannter Rebound-Effekt.

So genannte *Business Intelligence Systeme* unterstützen das Management von Unternehmen, Geschäftsentscheidungen zu treffen. Häufig werden dabei KI-Techniken eingesetzt, z. B. die Sentiment-Analyse von social media-Kanälen. Chancen sind das frühzeitige Erkennen, bessere Verständnis und schnelle Reagieren auf Kundenmeinungen – sowohl Wünsche als auch Kritik. Aber darin liegen auch Risiken. Werden Kunden wirklich oder nur vermeintlich besser verstanden? Es besteht die Gefahr einer Schein-Objektivität beziehungsweise Zahlengläubigkeit, die echter Einsicht entgegensteht. Insbesondere sind solche Analysen nur bedingt für strategische Zukunftsplanung geeignet, da sie nur die Vergangenheit abbilden können. Außerdem wächst mit der Bedeutung von solchen automatisierten Analysen die Gefahr der absichtlichen Manipulation, z. B. durch fake news, welche von Computerprogrammen (bots) generiert werden.

Beim *Hochfrequenzhandel* werden umfangreiche Börsentransaktionen automatisch mittels KI-Anwendungen durchgeführt. Chancen bestehen für einzelne Unternehmen hier auf zusätzliche Gewinne durch Ausnutzen von kurzlebigen Kursschwankungen. Für den Markt und damit die gesamte Wirtschaft bestehen aber auch erhebliche Risiken durch nicht kalkulierbare Rückkopplungseffekte. So können einzelne Fehlentscheidungen von vielen anderen Systemen wiederholt werden; ein gegenseitiges Aufschaukeln kann nicht vorhersehbare Folgen mit sich bringen.

Robotik wird intensiv in der industriellen Fertigung eingesetzt. Die Chancen liegen sowohl in einer gesteigerten Effizienz als auch in einer höheren Präzision. Aber grundsätzlich bedingt eine gesteigerte Effizienz das Risiko einer geringeren Stabilität. So birgt hocheffiziente just in time-Produktion mit minimierter Lagerhaltung die Gefahr eines Zusammenbruchs der Lieferketten, wie es bei der Corona-Pandemie der Fall war. Auch hier gilt, dass komplette Vernetzung – die Grundlage für die hohe Automatisierung – gleichzeitig Einfallstore für umfassende Cyberangriffe bietet.

Überwachungssysteme setzen zunehmend KI-Methoden ein, z. B. Systeme zur Gesichtserkennung an Flughäfen. Chancen solcher Systeme sind erhöhte Sicherheit, auch bei der Detektion von potenziellen Straftätern, die weltweit agieren. Damit verbunden ist aber auch das Risiko der ‚falschen Sicherheit', da solche Erkennungssysteme nicht perfekt sind und auf vielfältige Weise manipuliert werden können.

Komplette Vernetzung bietet Einfallstore für Cyberangriffe.

Schließlich kann KI vielfältig für *militärische Anwendungen* eingesetzt werden, z. B. in Form autonomer Gefechtsdrohnen. Die Chance besteht darin, dass Länder sich besser verteidigen und ihre eigenen Soldaten besser schützen können. Es gibt aber auch enorme Risiken. Zum einen wird das Töten von Menschen vereinfacht, was durch die Verbreitung jener autonomer Waffen auf alle Staaten zurückfallen kann. Aus ethischer Sicht besonders zu kritisieren ist, die Entscheidung über Leben und Tod Maschinen zu überlassen. Die Gefahr besteht, dass technische Möglichkeiten neue Realitäten hervorbringen; fangen die einen an, so müssen die anderen nachziehen.

Beiträge zum Schwerpunktthema

Zukunftsvorstellungen aus der Vergangenheit gehen Christian Vater und Eckhard Geitz auf den Grund und analysieren in ihrem Beitrag, wie in früheren Berichten der Enquete-Kommissionen des Deutschen Bundestags über künstliche Intelligenz und Digitalisierung gedacht wurde, welche Schwerpunkte gesetzt wurden und wie diese ausgestaltet wurden. Heutige Debatten zu Technik- und KI-Zukünften binden sie an Technikvergangenheiten an, stellen Querbezüge über mehrere Jahrzehnte her und arbeiten Bedeutungen früher Formatierungen heutiger Konzeptverständnisse heraus. Wie KI heute angelegt und ausgerichtet wird, ist oft in solchen Kartierungen und Präsentationsformen der Enquete-Kommissionen begründet, die vor vielen Jahren bereits Vorschläge zum Umgang mit künstlicher Intelligenz im politischen Raum gemacht haben. Die Ausprägung heutiger Debatten und heutige Entscheidungen gehen somit mitunter auf Prägungen zurück, die in frühen Dokumenten angelegt wurden. Heute wirkmächtig und analytisch aufschlussreich sind also technische Zukünfte der Vergangenheit. Heutige Darstellungsformen von KI prägen dementsprechend auch zukünftige Wirkungen und Entwicklungen.

Die zunehmende Verbreitung von Sprachassistenten, Chatbots und anderen sprachverarbeitenden Programmen nimmt Sascha Dickel zum Anlass, kommunizierende Technik kritisch zu beleuchten – vor allem im Hinblick auf menschliche Deutungen der ihr zugeschriebenen Autonomie, Kreativität und (In-)Transparenz. Mithin stellt sich der Bedarf nach Identifizierbarkeit des Menschen besonders in hybriden Kommunikationsensembles moderner mediatisierter Gesellschaften, in denen maßgeschneiderte KI-Systeme bis dato genuin menschliche Rollen einnehmen. Der regelmäßige Umgang mit entsprechenden künstlichen Agenten lässt den Schluss zu, dass die Anwender*innen (und Gestalter*innen) diesen Artefakten – zumindest implizit – kontextuelle Sinnverständnisse zuschreiben. In die Zukunft gedacht, verschwimmt die bislang kulturell bedeutsame Grenze zwischen dem Menschlichen und dem Künstlichen in der allgemeinen Wahrnehmung oder sie wird durch eskalierende Turing-Tests immer wieder nachzuschärfen sein. Der Autor schlägt daher zur Klärung dessen, was wir von ‚sprechender Technik' erwarten dürfen, einen hermeneutischen Ansatz des ‚Verstehens des Verstehens von Technik' vor, der die Gestaltung, Wahrnehmung und Praxis künstlicher Kommunikation systematisch untersuchen soll.

Bernd Beckert behandelt das Thema vertrauenswürdiger KI. Es gibt bereits eine ganze Reihe praktischer Leitfäden für die Implementation vertrauenswürdiger KI-Anwendungen. Bei der Betrachtung von Umsetzungsprojekten zeigt sich jedoch durchweg ein noch geringer Konkretisierungsgrad. Was können die Gründe für das Umsetzungsdefizit sein? Bernd Beckert führt drei Gründe an: time to market-Überlegungen seitens 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 dass deshalb spezifische Vorkehrungen notwendig sind.

Stefan Strauß widmet sich der kritischen KI-Kompetenz für die konstruktive Nutzung von KI. Ziel seines Beitrags ist es, Problembewusstsein und kritische KI-Kompetenz zu stärken. Gesellschaftlicher Diskurs sowie Forschung und Entwicklung zu den Risiken künstlicher Intelligenz sind aus seiner Sicht oft einseitig: Entweder mit Fokus auf wichtigen, aber praxisfernen ethischen Aspekten, oder auf technokratischen Ansätzen, um gesellschaftliche Probleme nur mit Technologie zu lösen.

Es braucht jedoch auch praktikable, problemorientierte Perspektiven zur Analyse KI-bedingter Risiken. Im Kern sind Diskrepanzen zwischen Systemverhalten und Nutzungspraktiken in KI-automatisierten Anwendungskontexten zentrale Auslöser für Bias und gesellschaftliche Risiken. Stefan Strauß formuliert in dem Zusammenhang ein zentrales Meta-Risiko von KI-Systemen: Verzerrungen (automation bias). Dafür stellt er einen analytischen Rahmen vor mit Indikatoren zur Erkennung von Verzerrungen in KI-Anwendungen.

Eine besondere Aufgabe in der Gestaltung von IT-Systemen kommt jenen Führungskräften zu, die KI-Implementierungen und Einführungen verantworten und aktiv steuern. Ihr Einsatz in der Vermittlung von Zielklarheit und Erwartbarem inklusive der erwarteten Veränderungen sind nach Marie Jung und Jörg von Garrel wesentliche Einflussfaktoren für die Vertrauensbildung und die Unterstützung von Handlungsakzeptanz bei den Mitarbeitenden. Eine in diesem Sinne personalfreundliche Implementierung von KI-Systemen ist der Schwerpunkt ihres Beitrags, in dem sie anhand eines in ihrer Forschung entwickelten Modells aufzeigen, wie Vertrauensbildung und Akzeptanz bereits in frühen Entwicklungs- und Einführungsphasen angelegt und gefördert werden und somit zu einer erhöhten Handlungsakzeptanz beitragen können.

Angesichts des komplexen und unübersichtlichen Rechtsrahmens für Datenschutz, Dateneigentum und Urheberrecht fragt sich Thomas Wilmer, ob rechtssichere Rahmenbedingungen für neue KI-Entwicklungen weiterhin durch klassische, politisch veranlasste Regulierungsansätze formuliert werden sollen oder ob hier niederschwellige Verfahrensweisen womöglich besser geeignet sind. Dabei tangieren die spezifischen Datenbedarfe und -angebote vieler anwendungsorientierter KI-Entwicklungen direkt oder zumindest indirekt auch persönliche Interessen Einzelner und sind insofern von rechtlicher Relevanz. Dies betrifft insbesondere die Bereiche der Arbeitswelt und der medizinischen Forschung und Versorgung. In diesen sensiblen Bereichen sind Transparenzgebote und Grenzen für den angemessenen Einsatz von KI-Systemen durch geeignete Regulierungen zu formulieren. Entsprechende Maßnahmen sollen der Datensouveränität der Betroffenen und der Rechtssicherheit der Systementwickler und -betreiber dienen. Diese Aufgabe ist angesichts der unterschiedlichen Geschwindigkeiten heutiger technischer und legislativer Entwicklungen und der Heterogenität der Einsatzgebiete von KI eine große Herausforderung für die Verantwortlichen. Der Autor schlägt daher bestimmte vertragliche bzw. Control by Design-Lösungen zur Schließung von Regelungslücken im Bereich der KI vor: Eine pragmatische Lösung für zulässige Datenauswertungen und -übertragungen durch KI-Systeme wäre die Formulierung privatrechtlicher Klauseln in entsprechenden Vereinbarungen mit den Beteiligten (z. B. in Form von AGBs), die einen Rahmen für den Verkehr sowohl personenbezogener als auch anonymer Daten setzen. Diese Lösung könnte gegebenenfalls auch eine (finanzielle) Anreizregelung für die Betroffenen beinhalten. Ein ergänzender, vorgelagerter Ansatz greift die Idee des Privacy by Design der Europäischen Datenschutz-Grundverordnung auf. Im Unterschied dazu soll der hier vorgeschlagene Control by Design-Ansatz die Souveränität auch über anonymisierte Daten bis hin zu entsprechenden Haftungsfragen regeln. Konkret könnte dies z. B. durch transparente und anerkannte Instrumente der informierten Einwilligung (Cookies etc.) unterstützt werden.

KI hat ein erhebliches Nutzungspotenzial in Forschung und Entwicklung – insbesondere auch in den Lebenswissenschaften und in der Biomedizin. Entsprechende Anwendungsbereiche zeigen sich in Medikamentenentwicklung, klinischer Praxis und vorgelagerter humangenetischer Forschung. Gleichwohl stehen diesen Potenzialen auch zahlreiche Herausforderungen entgegen, die besondere Handlungsnotwendigkeiten nach sich ziehen. Der Frage, wie die Entwicklung von KI mit der sich ebenfalls rasch entwickelnden Humangenomik konvergiert und welche praktische Bedeutung diese Konvergenz für die biomedizinische Praxis hat, haben sich Reinhard Heil und sechs weitere Ko-Autor*innen gewidmet. Auf Basis einer umfangreichen Literaturanalyse untersucht das Team, inwieweit der korrelative Zugang von KI zu neuem biomedizinischem Wissen dem Bedarf an begründbarem Wissen z. B. für die Diagnose und Therapie von Krankheiten gerecht wird und welche Fragen sich hieraus für Handlungsebene ergeben. Die Autor*innen heben dabei u. a. folgende ungelöste Probleme einer KI-unterstützten Forschung hervor: fragliche Qualität des Wissens und dessen Interpretation sowie dessen Erklärbarkeit, Evaluierung und Anerkennung. Auf anderer Ebene kommen noch ungeklärte Folgen von KI-assistierter Forschung für Wissenschaft, Wirtschaft und Gesellschaft dazu – auch hinsichtlich ethischer Herausforderungen sowie Fragen der adäquaten Gestaltung entsprechender Forschungs- und Dateninfrastrukturen und damit verknüpfter Regulierungsnotwendigkeiten (s. o.). Angesichts dieser Gemengelage empfehlen Heil et al. vertiefende Analysen, breite gesellschaftliche Dialoge zu kritischen Einzelproblemen sowie explorative Politikansätze, um einschlägige Forschungsprogramme und Regulierungen in allgemein wünschbarer Weise zu verbessern.

KI-Anwendungen haben bereits heute große Auswirkungen auf Gesellschaft, Wirtschaft und Wissenschaft und dies wird noch zunehmen. KI-Anwendungen richtig einzuschätzen und verantwortlich zu gestalten ist eine wichtige Aufgabe. Dieses Themenheft soll dazu einen Beitrag leisten.

Literatur

Allen, Frederick (2001): The myth of artificial intelligence. In: American Heritage 52 (1). Online verfügbar unter https://www.americanheritage.com/myth-artificial-intelligence, zuletzt geprüft am 04. 11. 2021.

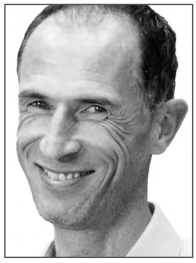
BMBF – Bundesministerium für Bildung und Forschung (2021): Nationale Strategie für Künstliche Intelligenz. Online verfügbar unter https://www.ki-strategie-deutschland.de, zuletzt geprüft am 04. 11. 2021.

Collingridge, David (1982): The Social Control of Technology. London: Pinter.

Gethmann, Carl Friedrich et al. (2021): Künstliche Intelligenz in der Forschung. Neue Möglichkeiten und Herausforderungen für die Wissenschaft. Cham: Springer Nature (in Druck).

Kurzweil, Ray (2015): Menschheit 2.0. Die Singularität naht. Berlin: Lola Books.

Marr, Bernard (2018): The key definitions of artificial intelligence (AI) that explain its importance. In: Forbes, 04. 02. 2018. Online verfügbar unter https://www.forbes.com/sites/bernardmarr/2018/02/14/the-key-definitions-of-artificial-intelligence-ai-that-explain-its-importance/#54d0f59a4f5d, zuletzt geprüft am 04. 11. 2021.



PROF. DR. BERNHARD G. HUMM
ist Informatik-Professor an der Hochschule Darmstadt und geschäftsführender Direktor des Instituts für Angewandte Informatik Darmstadt (aIDa). Sein Forschungsschwerpunkt ist Angewandte KI. Er führt nationale und internationale KI-Forschungsprojekte im Industrie- und Hochschulumfeld durch und publiziert regelmäßig.

PROF. DR. JAN C. SCHMIDT
ist promovierter Physiker und habilitierter Philosoph. Seit 2008 ist er Professor für Wissenschafts- und Technikphilosophie an der Hochschule Darmstadt. Schmidt ist Mitglied verschiedener Beiräte und Kuratorien, etwa dem Transdisziplinaritätsbeirat der Schweizerischen Akademie der Wissenschaften oder dem Konvent der ev. Akademie Frankfurt.

DR. STEPHAN LINGNER
verantwortet seit vielen Jahren den Forschungsbereich Technology Assessment am Institut für qualifizierende Innovationsforschung und -beratung GmbH (IQIB) in Bad Neuenahr-Ahrweiler. Er ist zudem Mitglied im Koordinations-Team des Netzwerks TA (NTA) sowie im Wissenschaftlichen Beirat der Zeitschrift für Technikfolgenabschätzung in Theorie und Praxis (TATuP).

PROF. DR. KARSTEN WENDLAND
ist Informatiker, Humanwissenschaftler und Technikfolgenabschätzer. An verschiedenen Universitäten und Hochschulen lehrt er in den Grenz- und Überlappgebieten von Technik- und Humanwissenschaften und forscht zu Sonderthemen im Feld Digitale Technologien und gesellschaftlicher Wandel am Institut für Technikfolgenabschätzung und Systemanalyse (ITAS).

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.*

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.

RESEARCH ARTICLE

Wenn die Technik sprechen lernt

Künstliche Kommunikation als kulturelle Herausforderung mediatisierter Gesellschaften

Sascha Dickel, *Institut für Soziologie, Johannes Gutenberg-Universität Mainz, Saarstr. 21, 55122 Mainz, DE*
(dickel@uni-mainz.de) *0000-0003-3620-2934*

Zusammenfassung • Wir scheinen am Anfang einer Ära kommunizierender Technik zu stehen. Amazons Sprachassistenzsystem Alexa ist in Millionen von privaten Haushalten eingezogen, Chatbots gehören heute zu den Standardanwendungen im Kundenservice und der Einsatz von Algorithmen zur Erzeugung von Texten ist in die Praxis professionellen Publizierens integriert worden. An kommunizierenden Maschinen zeigen sich kontroverse Deutungen von Autonomie, Kreativität, Komplexität und Opazität von Mensch und Maschine wie in einem Brennglas. Der Beitrag argumentiert, dass kommunizierende Technik eine eigenständige kulturelle Herausforderung für mediatisierte Gesellschaften darstellt, der sich eine reflexive Technikfolgenabschätzung hermeneutisch zu stellen hat.

When technology learns to speak. *Artificial communication as a cultural challenge for mediatized societies*

Abstract • *We are at the dawn of an era of communicating technology. Amazon's Alexa voice assistant system has entered millions of private households, chatbots have become a standard application in customer service, and the use of algorithms to generate texts has been integrated into the practice of professional publishing. Communicating machines reveal controversial interpretations of autonomy, creativity, complexity, and opacity of humans and machines. The article argues that artificial communication represents a cultural challenge for mediatized societies that reflexive technology assessment should address hermeneutically.*

Keywords • *artificial intelligence, artificial communication, hermeneutics, social theory*

https://doi.org/10.14512/tatup.30.3.23
Received: Jun. 09, 2021; revised version accepted: Oct. 13, 2021; published online: Dec. 20, 2021 (peer review)

Künstliche Kommunikation

Im gesellschaftlichen Diskurs der Digitalisierung spielt der Topos der künstlichen Intelligenz eine herausgehobene Rolle. Dieser Artikel nimmt diesbezüglich eine Problemverschiebung vor, indem er sich einem Phänomen annähert, das sich im Windschatten des gesellschaftlichen Diskurses um künstliche Intelligenz (KI) entfaltet und das Verhältnis von Mensch und Technik fundamental rekonfigurieren könnte. Es handelt sich um das Phänomen sprechender und schreibender Maschinen und damit um die Frage nach den Möglichkeiten und Grenzen *künstlicher Kommunikation*. Dieser Begriff wurde von Elena Esposito (2017, S. 261) geprägt: „By artificial communication I mean communication that involves an entity, the algorithm, which has been built and programmed by someone to act as a communication partner".

Bereits Alan Turing hat die Frage „Can machines think?" (Turing 1950, S. 433) als Problem der Kommunikationsfähigkeit von Maschinen reformuliert. Anstelle einer wie auch immer gearteten definitorischen Bestimmung oder Messung von Intelligenz, wird durch den (später sogenannten) Turing-Test geprüft, ob Maschinen in einer spezifisch formatierten Interaktionssituation von Menschen nicht mehr treffsicher zu unterscheiden sind. Die populärste Variante des Tests[1] basiert auf der Kommunikation eines Menschen (der ‚Richter:in') mit zwei weiteren Entitäten. Eine davon ist ein Mensch, die andere eine Maschine. Die Kommunikation findet dabei rein textbasiert, also explizit

1 An der Stelle sei angemerkt, dass die ursprüngliche gedankenexperimentelle Versuchsanordnung von Turing (1950) ein:e Richter:in, einen Mann und eine Frau vorsah. Die Aufgabe der Richter:in bestand darin, herausfinden, welcher der beiden anderen Akteure Mann oder Frau war. Turing fragte dann, was passieren würde, wenn der Mann durch eine Maschine ersetzt würde. Mittlerweile hat sich die oben skizzierte Form gleichwohl als „Standardvariante" des Tests etabliert.

ohne Sicht- und Hörkontakt, statt. Die Aufgabe der Richter:in besteht nun darin, zu entscheiden, welcher der beiden Interaktionspartner:innen ein Mensch ist und bei welchem es sich um eine Maschine handelt. Dabei sind alle möglichen Fragen und Antworten legitim. Die Aufgabe von Maschine (und Mensch!) besteht darin, die Richter:in davon zu überzeugen, ein Mensch zu sein. Je besser es Maschinen gelingt, Humanität vorzutäuschen, desto besser wird ihr Abschneiden im Test bewertet. Die ‚Intelligenz' von Maschinen wird beim Turing-Test also nicht an bestimmten Merkmalen festgemacht, die einem Programm selbst zu eigen sind und die z. B. mit einem Intelligenztest erfassbar wären, sondern an der maschinellen *Bewährung als Kommunikationspartner:in.*

Die Intelligenz von Maschinen wird beim Turing-Test nicht an bestimmten Merkmalen festgemacht, die einem Programm selbst zu eigen sind, sondern an der maschinellen Bewährung als Kommunikationspartner:in.

Der Test sollte nicht lange Gedankenexperiment bleiben. Bereits das 1966 von Joseph Weizenbaum entwickelte Programm ELIZA erregte Aufmerksamkeit, da es durch ein limitiertes, aber rollenkonformes (nämlich therapeutisch interpretierbares) Antwortverhalten als ‚echte:r' Interaktionspartner:in wahrgenommen werden konnte. Ähnlich wie beim Turing-Test war hier die Fähigkeit des Programms, sich an einer Konversation in natürlicher Sprache beteiligen zu können, wichtiger als die interne Komplexität des Programms (Suchman 2007, S. 47 ff.).

Insgesamt blieb die künstliche Nachbildung menschlichen Kommunikationsverhaltens ein stets wiederkehrendes Thema technikwissenschaftlicher Forschung und ihrer kulturellen Reflexion. Auch und gerade in der Science-Fiction wurden kommunikationsfähige Maschinen in zahlreichen Varianten portraitiert. Eine ganze Reihe von zeitgenössischen Filmen (etwa: Her, Ex Machina, Blade Runner 2049) präsentiert Konstellationen, in denen Menschen und Maschinen miteinander sprechen und dabei nicht zuletzt die Art und Weise ihrer sozialen Beziehung reflektieren.

Mittlerweile ist die Kommunikation mit künstlichen Systemen zu einer *Alltagserfahrung* geworden, die über spezifische Kontexte technikwissenschaftlicher Erprobung und kultureller Reflexion weit hinausgeht: Sprachassistenzsysteme wie Alexa und Siri bevölkern Millionen von privaten Haushalten und Hosentaschen, Chatbots gehören zur Standardanwendung im Kundenservice und der Einsatz von Algorithmen zur Erzeugung von Texten wird bereits in vielfältige Praktiken professionellen Publizierens integriert (Guzman und Lewis 2020). Die Gesellschaft hat sich offenbar zunehmend daran gewöhnt, dass maschinelle Entitäten am Sozialen qua Kommunikation teilnehmen (Sieber 2019; Hepp 2020). Zugleich entwerfen Technologiefirmen Szenarien der Mensch-Maschine-Beziehung, in denen kommunizierende Systeme noch wesentlich selbstverständlicher in Alltag und Berufsleben integriert sind und dabei menschliche Positionen einnehmen (Dickel und Schmidt-Jüngst 2021). In einer Welt kommunizierender Maschinen werden akademische Streitfragen um die Autonomie, Kreativität, Komplexität und Opazität von Mensch und Maschine in alltagspraktische Problemstellungen transformiert.

Bereits vor zehn Jahren stellte Dirk Baecker (2011) in dieser Zeitschrift die Frage, ob die Gesellschaft sich zukünftig auf Maschinen als Kommunikationspartner:innen einstellen müsse. Seine Überlegungen waren seinerzeit noch weitgehend theoretisch-spekulativer Natur. Heute jedoch hat die Frage nach den Implikationen kommunizierender Technik eine unverkennbare empirische Relevanz und Brisanz gewonnen, welche nach Antworten insbesondere auch aus dem Feld der Technikfolgenabschätzung (TA) verlangt. Zugleich ist aber aktuell noch kaum klar, was in den kommenden Jahren und Jahrzehnten technisch möglich sein wird, welche Konzepte realisiert werden und welche Anwendungen Akzeptanz finden. In genau solchen Situationen einer Unsicherheit von Zukunft, stoßen sowohl prognostische als auch szenariobasierte Verfahren der TA an ihre Grenzen. Stattdessen bieten sich *hermeneutische* Reflexionen an, die soziotechnische Visionen und Optionen vor ihrem gesellschaftlichen Horizont zu verstehen versuchen (Grunwald 2014). TA wird damit stärker als bislang sinnverstehend und gesellschaftsdiagnostisch.

Ich schlage in diesem Beitrag vor, sozial- und gesellschaftstheoretisches Wissen als Ressource für eine hermeneutische TA fruchtbar zu machen. Dieser Rückgriff dient dem Verständnis, welche Aspekte von Kultur und Gesellschaft durch eine technologische Option in welcher Weise berührt werden. Für den Gegenstand schreibender und sprechender Maschinen soll vor diesem Hintergrund zunächst die Bedeutung von Kommunikation ausgeleuchtet und gesellschaftsdiagnostisch eingebettet werden. Diese Reflexion ist zugleich eine erste hermeneutische Annäherung, die klären soll, worin die grundsätzliche Herausforderung des Phänomens (jenseits spezifischer Anwendungen) liegt. Der Artikel entwirft darauf aufbauend ein Konzept, wie spezifischere hermeneutische Forschungsanstrengungen aussehen könnten, die für das Feld der TA Orientierungs- und Handlungswissen generieren könnten.

Kommunikationstheoretische Reflexion

Eine ganze Reihe soziologischer Ansätze betonen, dass Kommunikation eine wesentliche, wenn nicht gar die entscheidende Rolle für die Konstitution sozialer Ordnung einnimmt (Habermas 1981; Knoblauch 2017; Luhmann 1984). Gesellschaft reproduziert sich durch kommunikative Akte, die aneinander anschließen und sukzessive soziale Wirklichkeit hervorbringen. Niklas Luhmann hat dazu einen überaus schlanken Kommunikationsbegriff vorgeschlagen, der mit sehr wenigen Vorannahmen über die ontologischen Beschaffenheiten der an Kommunikation beteiligten Akteur:innen auskommt. Gerade dies macht den Begriff für Untersuchungen attraktiv, die sich der Bedeutung von Kommunikation mit nicht-menschlichen Entitäten in Zeiten der Digitalisierung zuwenden wollen (Esposito 2017). Mit Luhmann (1984) kann Kommunikation als dreigliedriger Selektionsprozess begriffen werden, der *Information* (was wird gesagt), *Mitteilung* (wie wird es gesagt) und *Verstehen* (dass etwas gesagt wurde) umfasst. Dabei sollte Luhmann zufolge keine dieser drei Selektionen als innerpsychischer Prozess aufgefasst werden. Vielmehr bildet Kommunikation eine Ordnung sui generis, in der prozesshaft bestimmt wird, was eigentlich von wem mitgeteilt wurde, ob es zu Verstehensproblemen gekommen ist und inwiefern etwas überhaupt als Kommunikation gemeint war.[2] Das hier explizierte Modell von Kommunikation geht davon aus, dass die Entitäten, die an Kommunikation beteiligt sind, wechselseitig intransparent sind und bleiben. Gerade weil diese Intransparenz nicht intersubjektiv gelöst werden kann, kommt Kommunikation in Gang. Autonomie, Handlungsfähigkeit und Bewusstsein – die im Diskurs um ‚künstliche Intelligenz' typischerweise als *Eigenschaften* von Akteuren begriffen werden – werden kommunikationstheoretisch als *Zurechnungen* sichtbar, die kommunikativ explizit gemacht, stillschweigend vorausgesetzt oder (kommunikativ!) problematisiert werden können.

In der modernen Gesellschaft (sprich: in ihrer kommunikativen Praxis!) geht man gemeinhin davon aus, dass nur Menschen als *vollwertige* Teilnehmer:innen an gesellschaftlicher Kommunikation anerkannt sind. Sie werden *personifiziert,* während andere Entitäten (Ahnen, Bäume, Tiere, Geister, etc.) häufiger objektiviert bzw. fiktionalisiert werden. Gesa Lindemann (2011) hat herausgearbeitet, dass diese kommunikative Sonderstellung des Menschen zwar in der Moderne institutionalisiert wurde, aber prinzipiell als kulturell kontingent zu betrachten ist. Es ist eben diese vermeintlicherweise herausgehobene Position von Menschen als exklusiv kommunizierenden Personen, die mit der Ankunft sprechender und schreibender Maschinen herausgefordert wird. Das führt zu kulturellen Herausforderungen.

Kulturelle Herausforderungen

Eine erste Frage, die sich stellt, ist die nach der *Identität* des kommunikativen Gegenübers in einer Zeit maschineller Autor:innenschaft. Seit mehreren Jahren wird die Befürchtung artikuliert, dass sogenannte ‚Social Bots' die öffentliche Meinung auf sozialen Medien beeinflussen (Leistert 2017). 2018 stellte Google öffentlichkeitswirksam ein intelligentes persönliches Assistenzsystem (die Erweiterung Duplex für den Google Assistant) vor, welches die Stimme und das mündliche Kommunikationsverhalten von Menschen täuschend echt reproduzieren konnte. 2020 zeigte sich die Fachwelt beeindruckt, da der von dem Unternehmen OpenAI entwickelte Generative Pre-trained Transformer 3 (ein Sprachverarbeitungsmodell, das auf Deep Learning basiert) in der Lage ist, vielfältige Formen von Textgattungen zu generieren. Simone Natale (2021) weist angesichts solcher Phänomene darauf hin, dass es immer schwieriger werden könnte, Menschen und Maschinen zu unterscheiden, wenn künstliche Systeme sprechen und schreiben lernen und damit Merkmale menschlicher Kommunikation zu imitieren beginnen. Wenn Turing-Tests aus dem Labor in den Alltag wandern, stellen sich offenkundig neue Probleme der Zurechnung von Kommunikation.

Wenn aber Maschinen nun auch Bücher schreiben, Musik veröffentlichen, Beziehungen aufbauen und Dialoge führen, dann fordert dies das Verhältnis von Mensch und Technik in tiefgreifenderer Weise heraus.

Kommunizierende Maschinen lassen *zweitens* einen Problemhorizont aufscheinen, der über die typischen zeitgenössischen Deutungen von KI als einer neuen industriellen Revolution hinausgeht, in der nun (auch) die geistige Arbeit automatisiert wird (Brynjolfsson und McAfee 2012). Die hoffnungsvoll konturierte Komplementärdiagnose zur Maschinisierung kognitiver Leistung ist nämlich, dass nun menschliche Arbeiten verbes-

2 So kann A etwa einen Ausruf von B als Mitteilung betrachten, B kann bestreiten, dass dies so gemeint war und z. B. darauf hinweisen, dass sie sich lediglich den Fuß gestoßen hätte – und schon sind beide in einen Prozess des Kommunizierens verstrickt – und dass obwohl B niemals die Absicht hatte, A etwas mitzuteilen. Sozialität (im Kontrast zu bloßen Verhalten) kommt somit immer dann in Gang, wenn Entitäten unterstellt wird, dass sie nicht nur *das Eine,* sondern auch *etwas Anderes* hätten mitteilen oder verstehen können.

sert werden, die primär durch Kommunikation getragen werden: Care-Arbeit, Beziehungspflege, kreativ-schöpferische Tätigkeit, kontextsensibles Sprechen und Schreiben.

Wenn aber Maschinen nun auch Bücher schreiben, Musik veröffentlichen, Beziehungen aufbauen und Dialoge führen, dann fordert dies das Verhältnis von Mensch und Technik in tiefgreifenderer Weise heraus. Neben die Identitätsfrage tritt damit die Frage nach der *kommunikativen Substitution* des Menschen durch Maschinen (Sieber 2019). In den gegenwärtigen Einsatzgebieten kommunizierender Maschinen (vom automatisierten Journalismus über Service-Chatbots bis hin zu Sprachassistenzsystemen im Privatleben) ist diese Substitution gleichwohl noch auf spezifische Funktionsrollen fokussiert, die dezidiert an bestimmte Formen der Leistungserbringung gekoppelt sind (Lindemann 2011, S. 344). Das trifft auch dann zu, wenn kommunizierende Systeme explizit als Lebenspartner:innen oder Familienmitglieder vermarktet werden. Ein Beispiel für ersteres ist das japanische System Gatebox, ein Beispiel für letzteres der soziale Roboter Jibo. Dass wir Technik Funktionsrollen zuweisen, ist an sich nichts Besonderes. Brisant wird aber, wenn die dienstbare Technik zugleich durch kommunikative Zuschreibungen personifiziert wird. Nehmen wir das Beispiel von Alexa. Das System wird als moderne Servicekraft positioniert, die ihren Besitzer:innen im Umkehrschluss das Gefühl einer höheren sozialen Stellung vermitteln kann. Problematisch ist dabei nicht nur die durch Namen und Stimme vergeschlechtlichte Codierung der untergeordneten Funktionsrolle (Sontopski 2019; Dickel und Schmidt-Jüngst 2021). Befürchtet wird auch, dass die kommunikativen Muster (etwa von Befehl und Gehorsam), die mit maschinellen Kommunikationspartner:innen eingeübt werden, auf menschliche Kommunikationspartner:innen übertragen werden könnten.

Die dritte und letzte Herausforderung, die ich hier nennen möchte, ist zugleich die spekulativste: Werden wir geneigt sein, Maschinen zunehmend Autonomie, Bewusstsein und Verantwortlichkeit zuzuschreiben, wenn sie sich nicht nur in spezifisch formatierten Funktionskontexten, sondern vielfältigen lebenspraktischen Situationen kommunikativ bewähren? Aktuell sehen wir, dass die kommunikativen Fähigkeiten von Maschinen im Vergleich zu Menschen noch stark begrenzt sind. Die Interaktion mit einem Service-Chatbot wirkt typischerweise mechanisch und unflexibel. Und selbst bei anspruchsvolleren Systemen wie Alexa oder Siri, die eine mündliche Konversation simulieren sollen, wäre es kaum angebracht, von einem echten Dialog zu sprechen. Vielmehr handelt es sich um recht einfache Verkettungen von Fragen und Antworten (Sieber 2019). In den Zukunftsvisionen kommunizierender Maschinen werden solche Defizite freilich als technologische Hürden gerahmt, die prinzipiell überwindbar sind. Damit tritt auch der Traum von einer *Symmetrisierung* von Mensch und Maschine durch die Konstruktion kommunikativer Ebenbürtigkeit aktuell mit aller Macht in die Technoimagination unserer Gegenwart ein.

Mediatisierung als Bedingung maschineller Personifizierung

Ein hermeneutischer Blick auf technologische Entwürfe und Szenarien lenkt den Fokus auf die gesellschaftliche Gegenwart, in der sich oben skizzierte Herausforderungen als kulturelle Fragen überhaupt stellen. In welcher Gesellschaft also werden Fragen nach (1) der Identität von Mensch und Maschine, (2) der maschinellen Substitution von Menschen und schließlich (3) der Symmetrisierung von Mensch und Maschine eigentlich akut? Die zentrale These dieses Artikels ist, dass die *Mediatisierung von Kommunikation* (Krotz 2017) das entscheidende ‚Einfallstor' für die Personifizierung nicht-menschlicher Entitäten darstellt.

Kommunikation wird typischerweise immer noch oft von einem Paradigma körperlich anwesender Menschen verstanden, die ein mündliches Gespräch führen (Knorr Cetina et al. 2017). Doch ist diese Form der Kommunikation (nicht erst seit der Coronakrise) längst zu einer sehr spezifischen Spezialform geworden, neben die vielfältigste Formen des Kommunizierens auf Basis technischer Medien getreten sind (Dickel 2020). Neue medientechnologische Innovationen verändern die Strukturen des Sozialen nachhaltig, da sie die Formen des Kommunizierens verändern (Luhmann 1997, S. 249 ff.; Knoblauch 2017, S. 316 ff.; Baecker 2007). Durch mediale Infrastrukturen überschreitet das soziale Netz die Zeit- und Raumgrenzen der Interaktion unter körperlich Anwesenden: Autor:innen schreiben Texte für ein ihnen unbekanntes Publikum. Radio, Fernsehen und Internet schaffen neue technomediale Räume der Massenkommunikation. Aber auch die zwischenmenschliche Interaktion wird mediatisiert (Höflich 2016). Durch Medien wie Telefon, Chat und Instant Messaging-Programme werden synchrone Kommunikationskontexte medial angereichert und formatiert (Knorr Cetina et al. 2017). Mediale Techniken betreffen zunächst den *Mitteilungsaspekt* – also das ‚wie' – des Kommunizierens. Diese Veränderung von Mitteilungsmöglichkeiten verändert aber auch,

Die Mediatisierung von Kommunikation stellt das entscheidende ‚Einfallstor' für die Personifizierung nicht-menschlicher Entitäten dar.

was mitgeteilt wird und werden kann (*Informationsaspekt*), was überhaupt als Kommunikation gilt (*Verstehensaspekt*) – und welche Entitäten als Sender:innen und Empfänger:innen von Kommunikation infrage kommen.

Die Anwesenheit eines menschlichen Körpers ist längst nicht mehr nötig, damit Kommunikation als Kommunikation verstanden wird: Die Gesellschaft hat sich nicht nur daran gewöhnt, geschriebene Worte als Mitteilungen von Autor:innen zu interpretieren, sondern auch körperlose Stimmen aus Lautsprechern und Gesichter auf Bildschirmen als Repräsentationen von Anwesenden zu behandeln, denen Kommunikationsfähigkeit zugerechnet werden kann und bei denen Verstehen erwartet wird. Diese gesellschaftliche Einübung mediatisierter Kommunikation bereitet den Boden dafür, dass sich ‚hinter' dem medialen Interface nicht mehr zwingend menschliche Kommunikationsteilnehmende befinden müssen.

Die Anwesenheit eines menschlichen Körpers ist längst nicht mehr nötig, damit Kommunikation als Kommunikation verstanden wird.

Künstliche Kommunikation setzt an eben dieser Stelle an. Sie wird heute immer mehr zum expliziten Ziel der Gestaltung von Human-Machine-Interfaces (Sieber 2019; Hepp 2020). Das Aufkommen kommunizierender Maschinen lässt sich demgemäß als nächste Stufe einer „deep mediatization" (Hepp 2020, S. 1413) des gesellschaftlichen Lebens begreifen, in welcher sich die Bedeutung von Medien und Kommunikation selbst verändert: „Communicative AI departs from the historical role of media as mere channels of communication, since AI also acts as a producer of communication" (Natale 2021, S. 11).

Der Turing-Test lässt sich diesbezüglich weniger als Blaupause für maschinelle Denkfähigkeit lesen, sondern als kommunikations- und medientheoretisches Lehrstück.

- Der Test führt *erstens* vor, dass sich die Differenz von Mensch und Maschine durch eine kommunikative Praxis selbst situativ aktualisiert.
- Er demonstriert *zweitens*, dass die Kommunikation von Mensch und Maschine durch mediale Infrastrukturen geformt wird. Die gesamte Anordnung des Tests funktioniert nur unter mediatisierten Sonderbedingungen (hier: von Intransparenz und Schriftlichkeit).
- *Drittens* zeigt der Test, dass jede Kommunikation von Mensch und Maschine von gesellschaftlich institutionalisierten Deutungsmustern abhängt, die bestimmen, was überhaupt als typisch menschliches oder maschinelles Verhalten gilt.

Ob wir Technik Eigenschaften wie Autonomie, Reflexions- und Handlungsfähigkeit zubilligen, ist somit nicht nur eine Frage technischer Leistungsfähigkeit, sondern auch eine Frage medialer Settings und kultureller Praktiken, in die eine Technik eingebunden wird.

Jenseits des Turing-Tests: Eine hermeneutische Aufgabe für die TA

Dieser Artikel hat argumentiert, dass die Zuschreibung von nichtmenschlicher Kommunikationsfähigkeit als Mediatisierungseffekt zu begreifen ist und damit maßgeblich von den medialen Ökologien abhängt, in denen KI eingebettet ist. Vor diesem Hintergrund muss sich eine TA der künstlichen Kommunikation vor allem der Frage zuwenden, wie die Gestaltung von und der Umgang mit medialen Infrastrukturen und Interfaces die Deutung hervorbringt, dass technische Artefakte mit uns kommunizieren – dass sie uns Informationen mitteilen und unsere Worte verstehen.

Angesichts der kulturellen Unsicherheiten, die mit kommunizierenden Maschinen verbunden sind, soll somit zum Abschluss dieses Artikels skizziert werden, worin die kommende Aufgabe einer hermeneutischen TA zu diesem Problemfeld bestehen kann. Ich schlage vor, dass sie sich insbesondere als Hermeneutik zweiter Ordnung profilieren sollte, welche das *Verstehen des Verstehens* von Technik als ihren Gegenstand begreift. Ansätze dafür liefern methodische Herangehensweisen der sozialwissenschaftlichen Hermeneutik, die soziale Deutungen und Praktiken nicht einfach vor einem schon vermeintlich verstandenen gesellschaftlichen Kontext deuten, sondern durch eine minutiöse Rekonstruktion von Sinnstrukturen diesen Kontext erst zu erschließen versuchen (Sammet und Erhard 2018).

Eine Hermeneutik der mediatisierten Kommunikation mit Maschinen würde die Aufgabe von TA, Orientierungswissen bereitzustellen, mit dem Aufzeigen gesellschaftlicher Handlungsoptionen verzahnen. Über die allgemeine hermeneutische Reflexion grundlegender Konstitutionsbedingungen und Herausforderungen (welche dieser Artikel umreißen konnte) hinaus, ergibt sich so die Möglichkeit einer Hermeneutik der kommunikativen Praxis, die nicht selbst – vom akademischen Lehnstuhl aus – definiert, wie es um die Kommunikationsfähigkeit von künstlichen Systemen bestellt ist, sondern rekonstruktiv erschließt, was es bedeutet, wenn Technik zur sozialen Adresse wird.[3] Solche Untersuchungen können auf drei Ebenen ansetzen:

3 Erste Forschungsarbeiten, die in diese Richtung weisen, liegen bereits vor (Muhle 2016; Meister 2014; Mayer et al. 2020; Voss 2021).

- Um die *Gestaltung* von künstlicher Kommunikation zu erschließen, kann durch Expert:inneninterviews und teilnehmende Beobachtungen in Erfahrung gebracht werden, welche Erwartungen hinsichtlich der Kommunikation von Mensch und Maschine in das Design der Technik eingeschrieben werden.
- Um die *Praxis* künstlicher Kommunikation anhand der Interaktionen von Nutzer:innen mit gegenwärtig verfügbaren Anwendungen zu rekonstruieren, können sozialwissenschaftliche Verfahren wie Sequenz- und Konversationsanalyse zum Einsatz kommen. Mit diesen lässt sich zeigen, wie Menschen und Maschinen sich gegenseitig kommunikativ behandeln und welche sozialen Positionen und Rollen sie sich wechselseitig zuweisen.
- Um die *Imaginationen* künstlicher Kommunikation (etwa Werbevideos von Firmen oder Science-Fiction-Filme) zu verstehen, können öffentliche Darstellungen sprechender Maschinen zum Beispiel video- und bildanalytisch untersucht werden.

In welchen Bereichen lassen wir zu, dass Maschinen kommunikativ menschliche Rollen einnehmen?

Durch solche Untersuchungen können wir nicht nur etwas über die impliziten und expliziten Erwartungen erfahren, die in den entsprechenden Artefakten angelegt und eingeschrieben sind. Anhand der Analyse von gegenwärtigen (oder in Entwicklung befindlichen) Kommunikationsmaschinen können wir auch etwas über deren Möglichkeiten und Unzulänglichkeiten sowie ihre spezifischen Bedingungen und Begrenztheiten in vorausschauender Weise lernen. Ergänzend dazu führen uns Imaginationen der Kommunikation mit Maschinen vor, wie solche Interaktion im Idealzustand aussehen könnten, aber auch welche Hoffnungen und Befürchtungen mit ihnen verbunden werden. In jedem Fall erbringen solche hermeneutischen Analysen Distanzierungsgewinne, die über unsere gesellschaftliche Beziehung zu Technik aufklären. Sie eröffnen so Ansatzpunkte für öffentliche Dialoge. Denn das Wissen über die gegenwärtig real praktizierten Deutungen kommunizierender Maschinen und das Verstehen ihres gesellschaftlichen Verstehens, eröffnet zugleich Möglichkeiten, die Arten und Weisen zu verhandeln, wie Kommunikation im Zeitalter kommunizierender Technik gestaltet werden kann. Welche Ansprüche stellen wir an die Identifizierbarkeit von Menschen? In welchen Bereichen lassen wir zu, dass Maschinen kommunikativ menschliche Rollen einnehmen? Wie würde eine Gesellschaft aussehen, die Mensch und Maschine qua Kommunikation symmetrisiert?

Wie mit der Differenz von Mensch und Maschine in Zeiten künstlicher Kommunikation zukünftig umgegangen wird, wird von Konventionen kommunikativer Praxis ebenso abhängen, wie von dem Design medialer Interfaces und Infrastrukturen. Dabei lassen sich spekulative Extrempole unterscheiden: Zum einen wäre denkbar, dass die Konstellation des Turing-Test zu einer impliziten Alltagskonstellation wird und ständig neue Praktiken und Techniken in Stellung gebracht werden, um Menschen und Maschinen in tiefgreifend mediatisierten Umwelten weiterhin differenzieren zu können. Der Gegenpol wäre eine Gesellschaft, in der eben diese Relevanzsetzungen verschwinden, und die Differenz von Mensch und Maschine als einer kulturell bedeutsamen Unterscheidung de-institutionalisiert wird. Das Resultat wäre eine posthumane Ordnung, in der die einstmals scharfe, ‚natürliche' Differenz von Mensch und Maschine als kontingente kulturelle Konstruktion gelesen wird – ähnlich wie dies heute für verschiedene Binnendifferenzierungen des Humanen (wie Geschlechter oder Völker) in Anspruch genommen wird.[4]

Angabe von Finanzierungsquellen

Der vorliegende Forschungsartikel hat keine Förderung erhalten.

Literatur

Baecker, Dirk (2007): Studien zur nächsten Gesellschaft. Frankfurt am Main: Suhrkamp.

Baecker, Dirk (2011): Who qualifies for communication? A systems perspective on human and other possibly intelligent beings taking part in the next society. In: Technikfolgenabschätzung. Theorie und Praxis 20 (1), S. 17–26. https://doi.org/10.14512/tatup.20.1.17

Brynjolfsson, Erik; McAfee, Andrew (2012): Race against the machine. How the digital revolution is accelerating innovation, driving productivity, and irreversibly transforming employment and the economy. Lexington: Digital Frontier Press.

Dickel, Sascha (2020): Gesellschaft funktioniert auch ohne anwesende Körper. Die Krise der Interaktion und die Routinen mediatisierter Sozialität. In: Michael Volkmer und Karin Werner (Hg.): Die Corona-Gesellschaft. Analysen zur Lage und Perspektiven für die Zukunft. Bielefeld: transcript, S. 79–86. https://doi.org/10.14361/9783839454329-008

Dickel, Sascha; Schmidt-Jüngst, Miriam (2021): Gleiche Menschen, ungleiche Maschinen. Die Humandifferenzierung digitaler Assistenzsysteme und ihrer Nutzer:innen in der Werbung. In: Dilek Dizdar, Stefan Hirschauer, Johannes Paulmann und Gabriele Schabacher (Hg.): Humandifferenzierung. Disziplinäre Perspektiven und empirische Sondierungen. Weilerswist: Velbrück Wissenschaft, S. 342–367. https://doi.org/10.5771/9783748911364-342

4 Das Forschungsprogramm der Humandifferenzierung strebt an, 1) zwischenmenschliche Binnendifferenzierungen und 2) Unterscheidungen von Menschen und ihrer nicht-menschlichen Umwelt in ihrem Zusammenhang zu untersuchen. Vgl. dazu am Fall von Amazons Alexa Dickel und Schmidt-Jüngst 2021.

Esposito, Elena (2017): Artificial communication? The production of contingency by algorithms. In: Zeitschrift für Soziologie 46 (4), S. 249–265. https://doi.org/10.1515/zfsoz-2017-1014

Grunwald, Armin (2014): The hermeneutic side of responsible research and innovation. In: Journal of Responsible Innovation 1 (3), S. 274–291. https://doi.org/10.1080/23299460.2014.968437

Guzman, Andrea; Lewis, Seth (2019): Artificial intelligence and communication. A human-machine communication research agenda. In: New Media & Society 22 (1), S. 70–86. https://doi.org/10.1177/1461444819858691

Habermas, Jürgen (1981): Theorie des kommunikativen Handelns. Frankfurt am Main: Suhrkamp.

Hepp, Andreas (2020): Artificial companions, social bots and work bots. Communicative robots as research objects of media and communication studies. In: Media, Culture & Society 42 (7–8), S. 1410–1426. https://doi.org/10.1177/0163443720916412

Höflich, Joachim (2016): Der Mensch und seine Medien. Mediatisierte interpersonale Kommunikation. Eine Einführung. Wiesbaden: Springer VS. https://doi.org/10.1007/978-3-531-18683-2

Knoblauch, Hubert (2017): Die kommunikative Konstruktion der Wirklichkeit. Wiesbaden: Springer VS. https://doi.org/10.1007/978-3-658-15218-5

Knorr Cetina, Karin; Reichmann, Werner; Woermann, Niklas (2017): Dimensionen und Dynamiken synthetischer Gesellschaften. In: Friedrich Krotz, Cathrin Despotović und Merle-Marie Kruse (Hg.): Mediatisierung als Metaprozess. Transformationen, Formen der Entwicklung und die Generierung von Neuem. Wiesbaden: Springer VS, S. 35–57. https://doi.org/10.1007/978-3-658-16084-5_3

Krotz, Friedrich (2017): Mediatisierung. Ein Forschungskonzept. In: Friedrich Krotz, Cathrin Despotović und Merle-Marie Kruse (Hg.): Mediatisierung als Metaprozess. Transformationen, Formen der Entwicklung und die Generierung von Neuem. Wiesbaden: Springer VS, S. 13–32. https://doi.org/10.1007/978-3-658-16084-5_2

Leistert, Oliver (2017): Social Bots als algorithmische Piraten und als Boten einer techno-environmentalen Handlungskraft. In: Robert Seyfert und Jonathan Roberge (Hg.): Algorithmuskulturen. Über die rechnerische Konstruktion der Wirklichkeit. Bielefeld: transcript, S. 215–234. https://doi.org/10.14361/9783839438008-009

Lindemann, Gesa (2011): Die Akteure der funktional differenzierten Gesellschaft. In: Nico Lüdtke und Hironori Matsuzaki (Hg.): Akteur. Individuum. Subjekt. Wiesbaden: VS Verlag für Sozialwissenschaften, S. 329–350. https://doi.org/10.1007/978-3-531-93463-1_15

Luhmann, Niklas (1984): Soziale Systeme. Grundriß einer allgemeinen Theorie. Frankfurt am Main: Suhrkamp.

Luhmann, Niklas (1997): Die Gesellschaft der Gesellschaft. Frankfurt am Main: Suhrkamp.

Mayer, Henning; Muhle, Florian; Bock, Indra (2020): Whiteboxing MAX. Zur äußeren und inneren Interaktionsarchitektur eines virtuellen Agenten. In: Eckhard Geitz, Christian Vater und Silke Zimmer-Merkle (Hg.): Black Boxes. Versiegelungskontexte und Öffnungsversuche. Interdisziplinäre Perspektiven. Berlin: de Gruyter, S. 295–322. https://doi.org/10.1515/9783110701319-016

Meister, Martin (2014): When is a robot really social? An outline of the robot sociologicus. In: Michaela Pfadenhauer und Knud Böhle (Hg.): Of social robots and artificial companions. Contributions from the social sciences. STI Studies 10 (1), S. 107–134.

Muhle, Florian (2016): „Are you human?" Plädoyer für eine kommunikationstheoretische Fundierung interpretativer Forschung an den Grenzen des Sozialen. In: Forum Qualitative Sozialforschung 17 (1).

Natale, Simone (2021): Deceitful media. Artificial intelligence and social life after the Turing Test. New York: Oxford University Press. https://doi.org/10.1093/oso/9780190080365.001.0001

Sammet, Kornelia; Erhard, Franz (2018): Methodologische Grundlagen und praktische Verfahren der Sequenzanalyse. Eine didaktische Einführung. In: Franz Erhard und Kornelia Sammet (Hg.): Sequenzanalyse praktisch. Weinheim: Beltz Juventa, S. 15–71.

Sieber, Armin (2019): Dialogroboter. Wie Bots und künstliche Intelligenz Medien und Massenkommunikation verändern. Wiesbaden: Springer VS. https://doi.org/10.1007/978-3-658-24393-7

Sontopski, Natalie (2019): Hey Siri?! In: Kursbuch 55 (199), S. 62–75. https://doi.org/10.5771/0023-5652-2019-199-62

Suchman, Lucy (2007): Human-machine reconfigurations. Plans and situated actions. Cambridge, U. K.: Cambridge University Press. https://doi.org/10.1017/CBO9780511808418

Turing, Alan (1950): Computing machinery and intelligence. In: Mind LIX (236), S. 433–460. https://doi.org/10.1093/mind/LIX.236.433

Voss, Laura (2021): More than machines? Bielefeld: transcript. https://doi.org/10.14361/9783839455609

PROF. DR. SASCHA DICKEL
ist seit 2021 Professor für Mediensoziologie und Gesellschaftstheorie an der Johannes Gutenberg-Universität Mainz. Seine Arbeitsschwerpunkte sind Technikzukünfte, Mensch-Maschine-Beziehungen, Wissenschaftskommunikation und -partizipation.

OPEN ACCESS oekom

RESEARCH ARTICLE

Artificial intelligence in human genomics and biomedicine

Dynamics, potentials and challenges

Reinhard Heil, *Institute for Technology Assessment and Systems Analysis (ITAS), Karlsruhe Institute of Technology (KIT), Karlstr. 11, 76133 Karlsruhe, DE (reinhard.heil@kit.edu) 0000-0001-7265-7376*
Nils B. Heyen, *Fraunhofer Institute for Systems and Innovation Research ISI, Karlsruhe, DE (nils.heyen@isi.fraunhofer.de) 0000-0002-9354-1388*
Martina Baumann, *Institute for Technology Assessment and Systems Analysis (ITAS), Karlsruhe Institute of Technology (KIT), Karlsruhe, DE (martina.baumann@kit.edu)*
Bärbel Hüsing, *Fraunhofer Institute for Systems and Innovation Research ISI, Karlsruhe, DE (baerbel.huesing@isi.fraunhofer.de) 0000-0003-0277-0570*
Daniel Bachlechner, *Fraunhofer Austria Research GmbH, Innovation Center »Digital Transformation of Industry«, Wattens, AT (daniel.bachlechner@fraunhofer.at) 0000-0001-7726-9065*
Ulrich Schmoch, *Fraunhofer Institute for Systems and Innovation Research ISI, Karlsruhe, DE (ulrich.schmoch@isi.fraunhofer.de) 0000-0001-8615-5115*
Harald König, *Institute for Technology Assessment and Systems Analysis (ITAS), Karlsruhe Institute of Technology (KIT), Karlsruhe, DE (h.koenig@kit.edu) 0000-0002-0117-0939*

Abstract • The increasing availability of extensive and complex data has made human genomics and its applications in (bio)medicine an attractive domain for artificial intelligence (AI) in the form of advanced machine learning (ML) methods. These methods are linked not only to the hope of improving diagnosis and drug development. Rather, they may also advance key issues in biomedicine, e. g. understanding how individual differences in the human genome may cause specific traits or diseases. We analyze the increasing convergence of AI and genomics, the emergence of a corresponding innovation system, and how these associative AI methods relate to the need for causal knowledge in biomedical research and development (R&D) and in medical practice. Finally, we look at the opportunities and challenges for clinical practice and the implications for governance issues arising from this convergence.

Künstliche Intelligenz in der Humangenomik und Biomedizin.
Dynamiken, Potenziale und Herausforderungen

Zusammenfassung • *Die zunehmende Verfügbarkeit umfangreicher und komplexer Daten hat die Humangenomik und ihre Anwendungsbereiche in der (Bio-)Medizin zu einem attraktiven Bereich für künstliche Intelligenz (KI) vor allem in Form von fortgeschrittenen Methoden des maschinellen Lernens (ML) gemacht. Diese Methoden sind nicht nur mit der Hoffnung verbunden, Diagnosen und die Medikamentenentwicklung zu verbessern. Sie könnten auch darum, Kernthemen in der Biomedizin voranzubringen, z. B. zu verstehen, wie individuelle Unterschiede im menschlichen Genom bestimmte Merkmale oder Krankheiten verursachen können. Wir analysieren die zunehmende Konvergenz von KI und Genomik, das Entstehen eines entsprechenden Innovationssystems und wie diese assoziativen KI-Methoden mit dem Bedarf an kausalem Wissen in der biomedizinischen Forschung und Entwicklung und in der medizinischen Praxis zusammenhängen. Schließlich betrachten wir die Potenziale und Herausforderungen für die klinische Praxis und die sich aus dieser Konvergenz ergebenden Implikationen für Governance-Fragen.*

Keywords • *artificial intelligence, biomedicine, genomics, governance, knowledge*

https://doi.org/10.14512/tatup.30.3.30
Received: Jun. 17, 2021; revised version accepted: Oct. 08, 2021; published online: Dec. 20, 2021 (peer review)

Introduction

The increasing availability of extensive and complex data has made human genomics and its application areas in (bio)medicine an attractive domain for artificial intelligence (AI) in the form of advanced machine learning (ML) methods (Wainberg et al. 2018). The focus of interest is on sequence data of the human genome as well as on data of genes that are read (transcribed) or proteins that are produced in various body cells and organs. These can be combined with clinical data from biobanks or electronic patient records, among others. The use of ML in

human genomics and biomedicine is associated with the hope of obtaining answers to a key question in these domains, namely how individual differences or mutations in the genome cause specific traits or diseases. This would allow predictions about functional consequences of genetic differences, diagnoses, prognoses about therapy options or the development of new drugs at an unprecedented pace and scope.

With AI and human genomic research, two emerging techno-scientific domains appear to converge, both being linked to hopes and fears on issues such as new diagnosis and therapy options, health economics, autonomy, discrimination, privacy, or accountability – all of which will likely be judged against the background of different interests, values, or worldviews of people. To develop policies helping to align innovations with societal needs and expectations, (1) an understanding of scientific-technical potentials and challenges, and (2) (mutual) learning on perspectives through dialog activities on the technologies and possible applications involving various stakeholders and publics will be needed.

Given the nascent state of the convergence of both domains, in this article we mainly focus on the first step, by exploring and using evidence from the literature as well as perspectives from stakeholders involved in current scientific-technical developments. This work may inform discussions on policy issues as well as realistic scenarios in societal dialog activities in the required second step towards the further development of AI in human genomics according to societal expectations.

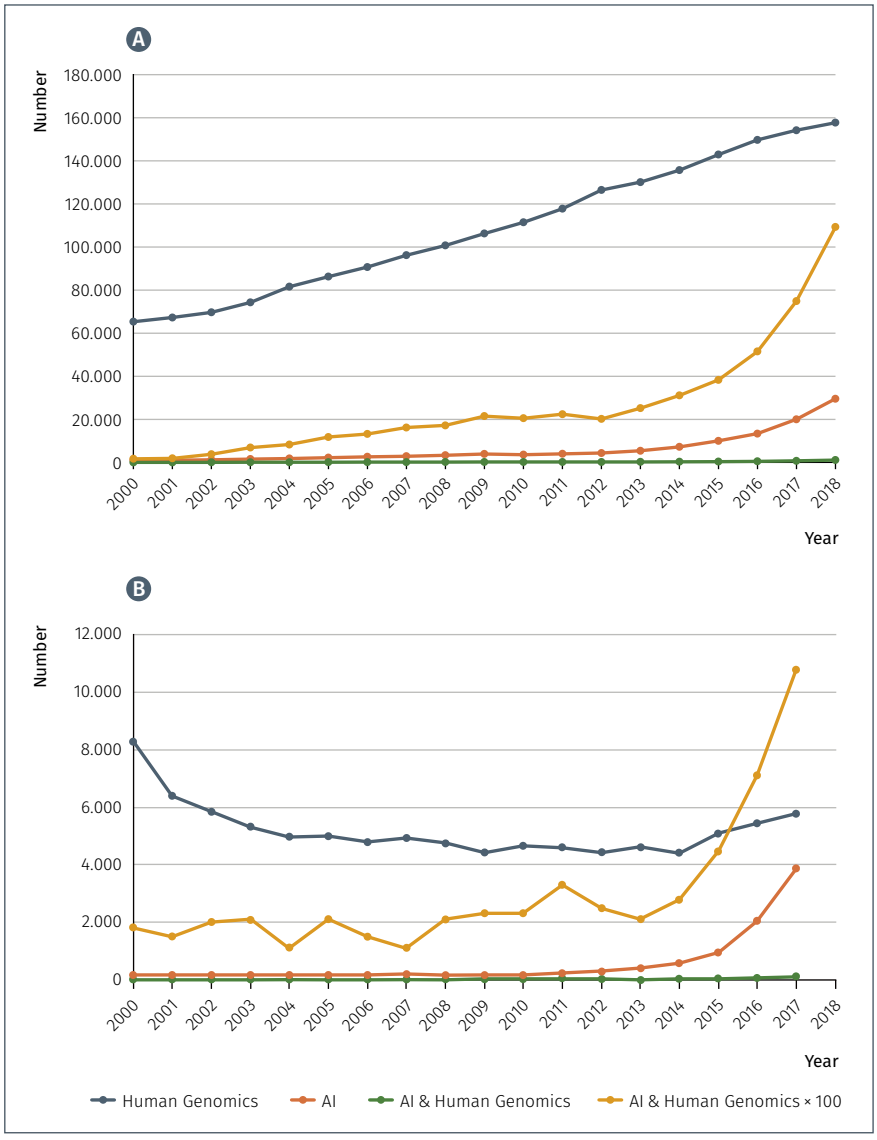


Fig. 1: Number of publications and patents worldwide over time.
A: Number of publications *(Source: Web of Science 2019. Authors' own calculations)*, **B:** Number of transnationally registered patents *(Source: World Patents Index 2019. Authors' own calculations).*

Number of publications and patents, respectively, in the domain 'AI in human genomics' (green curves), human genomics (blue) and AI (orange). To better visualize the dynamic development, publication and patent numbers of the green curves were multiplied by 100 (yellow curves).

Methods

As basis for identifying potentials and challenges, the current state of research as well as existing and emerging applications were examined and mapped by reviewing peer-reviewed scientific publications, conference proceedings, and patents (König 2020). In addition, the emerging innovation system at the intersection of AI and human genomics was analyzed through a mapping of international key actors, a review of international policy strategies, and a publication and patent analysis. The results of these steps were presented and discussed at a two-day workshop (held in November 2019 in Heidelberg) with twelve international experts from academic research, industry and the venture capital sector. Subsequently, the literature was evaluated with regard to ethical, social and regulatory challenges. These results informed a one-day workshop (held online in October 2020) with eight experts from Germany representing clinical research, genetic counseling, patient associations, medical and technical ethics, and jurisprudence. Finally, the results of all mentioned steps were reviewed and summarized for this article.

Results and Discussion

Innovation system analysis

Our analyses of the publication and patent statistics show that the use of AI-based methods in human genomics and biomedicine has boomed in recent years, both in academia and industry. Thus, the number of scientific publications and patents rapidly increased since 2014 (Figure 1 A/B; yellow curves). However, if one compares these activities to those in the domains of AI or of human genomics overall, it becomes clear that they still represent a relatively small niche with a limited number of key players and application areas (Figure 1 A/B; green curves). However, if this momentum continues, research and development (R&D) in human genomics will be heavily influenced by AI in the next years.

In the future, there could be a shift from scientific explanation based on experimentally validated causal relationships towards explanation largely derived from AI predictions.

Looking at different world regions, the publication analysis suggests that Europe as a whole plays a significant role in the global research domain (with the UK and Germany as the most prominent European countries) alongside the two dominant countries USA and China. In contrast, the statistics of transnational patents (suggesting a particular commercial relevance) show that the USA is well ahead of all other countries and regions in utilizing relevant research for commercial purposes.

The dynamics in global publication and patent activities at the intersection of AI and human genomics point to the emergence of a new innovation system. Our analyses suggest various factors and actors that have contributed to this development. These are in particular the widespread (also clinical) use of and demand for whole genome sequencing and corresponding platform technologies as well as the engagement of venture capital companies, big tech companies, pharma companies, and startups. The latter (especially in the USA) are significantly driving technological innovations for drug discovery and development (DKA 2019). In addition, numerous public research organizations, international research consortia (e. g., the International Cancer Genome Consortium) or local research clusters (such as Boston, MA, USA) are important drivers.

Potentials of AI for human genomics and biomedicine

Basic and translational research: Deep learning (DL), a form of ML which relies on software-simulated multiple layers of so-called artificial neurons (deep neural networks), is increasingly used to explore how individual differences in the human genome may cause traits or diseases. Several recent studies suggest that DL models, via analysis of large data sets, can be of decisive help in rapidly and comprehensively identifying putative causative genetic alterations and molecular mechanisms for diseases. These include common diseases such as neurological diseases (e. g., Alzheimer's disease, autism or schizophrenia), inflammatory bowel diseases (e. g., Crohn's disease) and diabetes (König et al. 2021; Zou et al. 2019). Similarly, potential pathogenic mutations that could facilitate diagnoses or prognoses have been identified for various cancer types and rare genetic diseases (Bailey et al. 2018; Brasil et al. 2019).

Drug development: ML methods are increasingly developed and used, especially by start-up companies, to make predictions of molecular properties (such as protein structures and interactions, or toxicity) from genomic data, which are used to identify drug candidates (virtual screening), to use existing drugs for new purposes (drug repurposing) or to design new drugs (Paul et al. 2021).

Clinical practice: ML has enabled new diagnostic methods by linking genomic data with clinical data (such as on disease progression or medical images). Such new methods include the diagnosis of rare genetic diseases in children by linking altered facial features, symptoms and genetic changes (Gurovich et al. 2019) or methods to analyze minute amounts of DNA and other tumor cell components in body fluids (liquid biopsies) (Heidrich et al. 2021). Another strand of development aims to improve risk predictions by combining and weighing very large numbers of individual genetic variants (genome-wide polygenic scores) for a range of important and common diseases to such an extent that they may be widely used clinically (Lello et al. 2019).

Challenges and fields of action

AI-based understanding, quality of knowledge, and potential implications: Although causality and causal modeling have become an active area of research in AI, currently established ML methods for analyzing large and complex data are still based on statistical modeling and do not reflect true causal relations but correlative associations (Schölkopf 2019). In contrast, current self-understanding of scientific explanation and associated quality criteria for (causal) knowledge in biomedical research rely on experimental intervention to reveal causal processes and interactions that lead to the event (e. g., a disease) to be explained (MacArthur et al. 2014; Soldner and Jaenisch 2018). Accordingly, rigorous validation of such correlation-based model predictions on putative gene functions or physiological pathways by interventions in molecular and cellular processes remains necessary. Due to the very large and increasing numbers of genetic

variants associated with complex traits and the complexity of implicated gene and protein networks, such validation is costly and dependent on the availability of suitable human cell models or model organisms.

In the future, there could thus be a shift from scientific explanation based on experimentally validated causal relationships towards explanation largely derived from AI predictions. This would not only challenge the current self-conception in basic research regarding scientific understanding and the quality of knowledge, but also pose challenges for application-oriented research. The latter is supported by retrospective studies of drug approvals, suggesting that demonstrating a causal genetic link between the drug target and the disease significantly increases the likelihood of successful drug development (King et al. 2019).

It remains unclear what influence explainability actually has on trust in and acceptance of AI-based systems.

While causal mechanistic accounts of understanding prevail in basic and translational biomedical research, in clinical practice and evidence-based medicine difference-making probabilistic concepts of causation are the centerpiece. Using randomized controlled trials as their most important tool, they usually only provide black-box causal claims about the (statistical) effectiveness of interventions in a studied population, without providing a mechanistic explanation (König et al. 2021). Therefore, and as the value of mechanistic knowledge is controversial among practitioners (Andersen 2012; Reiss and Ankeny 2016), the impact of a potential quality loss in causal-mechanistic knowledge through AI is much less clear. Yet causal knowledge on mechanisms can play a role in the interpretation of clinical trials (Andersen 2012) as well as in diagnosis when cases are rare or complex (Brush Jr. et al. 2017).

Research and data infrastructures as well as data governance: Large amounts of high-quality genomic and other 'omic' data as well as health data are essential for ML methods (Saunders et al. 2019; Wainberg et al. 2018). Harnessing possible benefits from these techniques for biomedicine would thus require the (further) development of large and diverse biobanks (Denny 2019) as well as international initiatives which link national genome and health data and allow as many researchers as possible to share and access data (Powell 2021; Saunders et al. 2019). This poses considerable challenges linked to sufficient data processing and storage capacity, broad implementation of common technical standards, such as the FAIR principles (Wilkinson et al. 2016), high data security, and solutions that enable data sovereignty (Phillips et al. 2020; Powell 2021; Saunders et al. 2019).

Although commercial or 'community' cloud computing services can in principle solve these challenges (Langmead and Nellore 2018), problems may arise regarding regulation and/or privacy aspects, e. g., if the cloud services are located outside a certain geographical area and data are subject to less strictly regulated access by security and law enforcement authorities (Dove et al. 2015; Kolata and Murphy 2018). Possible privacy issues that may lead to (re-)identification and discrimination risks of data donors also pose a major challenge beyond ethical and legal issues, because academic as well as industrial research relies on a sufficiently high representativeness of data sets. This requires that as many patients and study participants as possible make their data available. Ethnic groups in particular must not be underrepresented in the databases (Sirugo et al. 2019) in order to avoid biases and to generate benefits for as many people as possible.

In the area of private R & D, in addition to big tech companies, large pharmaceutical companies have increasingly invested in AI for diagnosis and drug development and have entered into numerous collaborations with innovative start-up companies. While data might become concentrated in pharmaceutical companies, start-ups are better placed to produce innovative approaches to using data, as they are usually more agile. They can take risks, pivot, focus on niche markets and be disruptive. Large companies often become sustainers as they chase quarterly results. They concentrate on incremental innovations that support their business models. Operating like a start-up through largely autonomous entities or collaborating with real start-ups is considered essential for new AI developments in the biomedical domain (DKA 2019). Large companies thus need to engage directly with start-ups by providing equity as well as access to resources such as technology and data. In Europe, however, the financing of such start-ups is a structural weakness due to various problems generally associated with the European venture capital landscape (DKA 2019).

Explainability, evaluation, and approval: Important ML methods for genomic medicine, especially DL models, have a distinct 'black box' character. They are difficult for humans to explain or interpret in terms of how and/or why a result is produced (Lipton 2018). This challenge is particularly severe for systems that are continuously learning and changing (Babic et al. 2019). The lack of explainability or interpretability is widely considered to be particularly important in the medical domain because of the high risks for human lives associated with potential errors and biases in models and data. Expectations are therefore high for so-called explainable AI systems (Arrieta et al. 2020). Despite the importance often attributed to explainability or interpretability, existing and proposed guidelines and regulations for AI in general, as well as for software as a medical device in particular (Ordish et al. 2019), currently lack clear standards for explainability or interpretability in both Europe and the USA. For

instance, in the USA, guidelines by the U. S. Food and Drug Administration (FDA) urge developers to provide information such as an "explanation of how the software works" (FDA 2019 b, 26), and physicians' ability to "independently review the basis for the recommendations" is considered important in determining whether software should be regulated (FDA 2019 a, 8). Similarly, the recently proposed EU Regulation on AI (AI Act) demands that high-risk AI systems are designed in a way "that their operation is sufficiently transparent to enable users to interpret the system's output and use it appropriately" (EC 2021, Article 13).

However, it remains largely unclear how information toward such transparency should look like in practice. In addition, the conceptual problem arises that for current ML systems it is only possible to explain how correlations and the predictions based on them are obtained, but not to draw causal conclusions (Pearl 2010). Thus, policy makers and regulators need to agree on more concrete information developers should provide with regard to the functionality of AI systems and on what role explainability can or should play for the approval of AI systems in comparison to complex, rigorous clinical trials (and possibly post-market monitoring).

Ethical and social implications: In the application of AI-based procedures in medical care, like diagnosis or treatment selection, trust and acceptability on the part of physicians and patients are seen as being of decisive importance (Arrieta et al. 2020; Kelly et al. 2019). Various EU policy documents, including the Ethics Guidelines for Trustworthy AI (HLEG 2019) and the European Commission's White Paper on AI (EC 2020), as well as the recent proposal for the EU's AI Act (EC 2021) aim at trustworthy AI. Transparency in the form of explainability or interpretability of AI systems (in the AI Act, in particular) is delineated as an important element to achieve this aim and the ethical use of AI.

However, given rather weak and contradictory empirical evidence, it remains unclear what influence explainability or interpretability actually have on trust in AI-based systems or their recommendations, compared to other factors, such as marketing, clinical trial data, or the regulatory environment. Studies on the diffusion or adoption of medical innovations also indicate that the adoption of applications is a complex social process (Azoulay 2002; Lublóy 2014). Thus, there is currently a lack of empirical evidence on how AI applications and their governance need to be designed in order to create and deserve sustainable trust and acceptance in medical AI systems. Before thinking about creating trust and acceptance, it is, of course, paramount that the AI system in question can or will actually improve the quality of care for specified patient groups. Trust is closely related to acceptance and both may be generated by explainability. However, since explainability cannot replace validation of AI systems with regard to medical outcomes and patient benefits by clinical trials, the generation of trust by 'plausible' explanations on how AI systems work just for the sake of pushing the diffusion of AI systems is highly problematic.

If risk predictions and disease progression prognosis, both in the direct-to-consumer and the regulated domain, should become available for a growing number of diseases and people, further challenges on the societal, health care system and individual level may arise. There would be a growing need for trained human genetic counselors who can help healthy and diseased individuals to make informed decisions (Heyen 2016), considering uncertainties arising from the potentially changed understanding of knowledge through AI systems. Already known social and psychological issues of health predictions or diagnoses based on genetics – such as societal pressure on individuals with regard to lifestyle related diseases or the right not to know (also of biological relatives) in case of a lack of therapeutic options (Voorwinden et al. 2020) – may be exacerbated if such AI systems are widely adopted. Health inequalities may be increased in case of a lack of financial support by health insurances for effectively health enhancing but costly genetic testing and AI-based diagnoses or prognoses. Research on these issues, especially by actively involving patients or the broader public is still scarce compared to the rapidly increasing technological possibilities.

The generation of trust by 'plausible' explanations on how AI systems work just for the sake of pushing the diffusion of AI systems is highly problematic.

Conclusion

In view of the epistemological, economical, technical, ethical and social challenges outlined above, as well as the current scarcity of evidence on how to best govern them, more research and efforts to experimental policy making are urgently needed. Given the complexity and wide scope of these challenges, broad societal debate and mutual learning by different forms of inclusive dialog activities, will be needed to improve research agendas and current regulatory proposals. These activities should involve stakeholders and publics – in order to harness the potentials and minimize risks for improved quality of care and life. By exploring and providing an overview of possible applications, actors and challenges, the article strives to help to identify and discuss most realistic scenarios in the needed societal dialogue.

Acknowledgement

Funding has been provided by the German Federal Ministry of Education and Research, Grant/Award Number: 16ITA201A/B.

References

Andersen, Holly (2012): Mechanisms. What are they evidence for in evidence-based medicine? In: Journal of evaluation in clinical practice 18 (5), pp. 992–999. https://doi.org/10.1111/j.1365-2753.2012.01906.x

Arrieta, Alejandro et al. (2020): Explainable artificial intelligence (XAI). Concepts, taxonomies, opportunities and challenges toward responsible AI. In: Information Fusion 58, pp. 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

Azoulay, Pierre (2002): Do pharmaceutical sales respond to scientific evidence? In: Journal of Economics & Management Strategy 11 (4), pp. 551–594. https://doi.org/10.1111/j.1430-9134.2002.00551.x

Babic, Boris; Gerke, Sara; Evgeniou, Theodoros; Cohen, I. Glenn (2019): Algorithms on regulatory lockdown in medicine. In: Science 366 (6470), pp. 1202–1204. https://doi.org/10.1126/science.aay9547

Bailey, Matthew et al. (2018): Comprehensive characterization of cancer driver genes and mutations. In: Cell 174 (4), pp. 1034–1035. https://doi.org/10.1016/j.cell.2018.07.034

Brasil, Sandra; Pascoal, Carlota; Francisco, Rita; Dos Reis Ferreira, Vanessa; Videira, Paula; Valadao, Goncalo (2019): Artificial intelligence (AI) in rare diseases. Is the future brighter? In: Genes 10 (12), pp. 1–24. https://doi.org/10.3390/genes10120978

Brush Jr., John; Sherbino, Jonathan; Norman, Geoffrey (2017): How expert clinicians intuitively recognize a medical diagnosis. In: The American journal of medicine 130 (6), pp. 629–634. https://doi.org/10.1016/j.amjmed.2017.01.045

Denny, Joshua et al. (2019): The 'All of Us' research program. In: New England Journal of Medicine 381 (7), pp. 668–676. https://doi.org/10.1056/NEJMsr1809937

DKA – Deep knowledge analytics "Pharma Divison" (2019): AI for drug discovery, biomarker development and advanced R & D landscape overview 2019/Q3. Available online at http://analytics.dkv.global/data/pdf/AI-for-DD-Q3-2019/AI-for-Drug-Discovery-Q3-2019-Teaser.pdf, last accessed on 05. 10. 2021.

Dove, Edward; Joly, Yann; Tassé, Anne-Marie; Knoppers, Bartha (2015): Genomic cloud computing. Legal and ethical points to consider. In: European Journal of Human Genetics 23 (10), pp. 1271–1278. https://doi.org/10.1038/ejhg.2014.196

EC – European commission (2020): White paper on artificial intelligence. A European approach to excellence and trust. Available online at https://ec.europa.eu/info/publications/white-paper-artificial-intelligence-european-approach-excellence-and-trust_de, last accessed on 28. 09. 2021.

EC – European commission (2021): Proposals for a regulation of the European Parliament and of the Council. Laying down harmonized rules on artificial intelligence (Artificial Intelligence Act) and amending certain union legislative acts. Available online at https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52021PC0206&from=EN, last accessed on 28. 10. 2021.

FDA – U. S. Department of Health and Human Services, Food and Drug Administration (2019 a): Clinical decision support software. Draft guidance for industry and food and drug administration staff. 27. 09. 2019. Available online at https://www.fda.gov/media/109618/download, last accessed on 28. 10. 2021.

FDA – U. S. Food and Drug Administration (2019 b): Developing a software precertification program. A working model. V1.0. Available online at https://www.fda.gov/media/119722/download, last accessed on 28. 10. 2021.

Gurovich, Yaron et al. (2019): Identifying facial phenotypes of genetic disorders using deep learning. In: Nature Medicine 25, pp. 60–64. https://doi.org/10.1038/s41591-018-0279-0

Heidrich, Isabel; Ačkar, Lucija; Parinanz, Mossahebi Mohammadi; Pantel, Klaus (2021): Liquid biopsies. Potential and challenges. In: International Journal of Cancer 148 (3), pp. 528–545. https://doi.org/10.1002/ijc.33217

Heyen, Nils (2016): Towards a technocratic biomedicine? In: Soziale Welt 67 (4), pp. 389–406. https://doi.org/10.5771/0038-6073-2016-4-389

HLEG – High-Level Expert Group on Artificial Intelligence (2019): Ethics guidelines for trustworthy AI. Available online at https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60419, last accessed on 28. 09. 2021.

Kelly, Christopher; Karthikesalingam, Alan; Suleyman, Mustafa; Corrado, Greg; King, Dominic (2019): Key challenges for delivering clinical impact with artificial intelligence. In: BMC medicine 17 (195), pp. 1–9. https://doi.org/10.1186/s12916-019-1426-2

King, Emily; Davis, Wade; Degner, Jacob (2019): Are drug targets with genetic support twice as likely to be approved? In: PLoS Genetics 15 (12), pp. 1–20. https://doi.org/10.1371/journal.pgen.1008489

Kolata, Gina; Murphy, Heather (2018): The golden state killer is tracked through a thicket of DNA, and experts shudder. In: The New York Times, 27. 04. 2018. Available online at https://www.nytimes.com/2018/04/27/health/dna-privacy-golden-state-killer-genealogy.html, last accessed on 05. 10. 2021.

König, Harald (2020): Artificial intelligence in human genomics. Mapping research and existing and emerging applications. Karlsruhe: Institute for Technology Assessment and Systems Analysis, pp. 1–9. https://doi.org/10.5445/IR/1000130769

König, Harald; Frank, Daniel; Baumann, Martina; Heil, Reinhard (2021): AI models and the future of genomic research and medicine. True sons of knowledge? In: Bioessays 43 (2100025), pp. 1–12. https://doi.org/10.1002/bies.202100025

Langmead, Ben; Nellore, Abhinav (2018): Cloud computing for genomic data analysis and collaboration. In: Nature Reviews Genetics 19 (4), pp. 208–219. https://doi.org/10.1038/nrg.2017.113

Lello, Louis; Raben, Timothy; Yong, Soke; Tellier, Laurent; Hsu, Stephen (2019): Genomic prediction of 16 complex disease risks including heart attack, diabetes, breast and prostate cancer. In: Scientific Reports 9 (15286), pp. 1–16. https://doi.org/10.1038/s41598-019-51258-x

Lipton, Zachary (2018): The mythos of model interpretability. In: Queue 16 (3), pp. 31–57. https://doi.org/10.1145/3236386.3241340

Lublóy, Ágnes (2014): Factors affecting the uptake of new medicines. A systematic literature review. In: BMC health services research 14 (469), pp. 1–25. https://doi.org/10.1186/1472-6963-14-469

MacArthur, Daniel et al. (2014): Guidelines for investigating causality of sequence variants in human disease. In: Nature 508, pp. 469–476. https://doi.org/10.1038/nature13127

Ordish, John; Murfet, Hannah; Hall, Allison (2019): Algorithms as medical devices. Cambridge, U. K.: PHG Foundation. Available online at https://www.phgfoundation.org/media/74/download/algorithms-as-medical-devices.pdf?v=1&inline=1, last accessed on 28. 09. 2021.

Paul, Debleena; Sanap, Gaurav; Shenoy, Snehal; Kalyane, Dnyaneshwar; Kalia, Kiran; Tekade, Rakesh (2021): Artificial intelligence in drug discovery and development. In: Drug discovery today 26 (1), pp. 80–93. https://doi.org/10.1016/j.drudis.2020.10.010

Pearl, Judea (2010): An introduction to causal inference. In: The International Journal of Biostatistics 6 (2), pp. 1–59. https://doi.org/10.2202/1557-4679.1203

Phillips, Mark et al. (2020): Genomics. Data sharing needs an international code of conduct. In: Nature 578 (7793), pp. 31–33. https://doi.org/10.1038/d41586-020-00082-9

Powell, Kendall (2021): The broken promise that undermines human genome research. In: Nature 590 (7845), pp. 198–201. https://doi.org/10.1038/d41586-021-00331-5

Reiss, Julian; Ankeny, Rachel (2016): Philosophy of medicine. In: The Stanford Encyclopedia of Philosophy. Available online at https://plato.stanford.edu/entries/medicine/, last accessed on 28. 09. 2021.

Saunders, Gary et al. (2019): Leveraging European infrastructures to access 1 million human genomes by 2022. In: Nature Reviews Genetics 20 (11), pp. 693–701. https://doi.org/10.1038/s41576-019-0156-9

Schölkopf, Bernhard (2019): Causality for Machine Learning. Available online at https://arxiv.org/abs/1911.10500, last accessed on 28. 09. 2021.

Sirugo, Giorgio; Williams, Scott; Tishkoff, Sarah (2019): The missing diversity in human genetic studies. In: Cell 177 (4), pp. 26–31. https://doi.org/10.1016/j.cell.2019.02.048

Soldner, Frank; Jaenisch, Rudolf (2018): Stem cells, genome editing, and the path to translational medicine. In: Cell 175 (3), pp. 615–632. https://doi.org/10.1016/j.cell.2018.09.010

Voorwinden, Jan et al. (2020): Cognitive and affective outcomes of genetic counselling in the Netherlands at group and individual level. A personalized approach seems necessary. In: European journal of human genetics 28 (9), pp. 1187–1195. https://doi.org/10.1038/s41431-020-0629-5

Wainberg, Michael; Merico, Daniele; Delong, Andrew; Frey, Brendan (2018): Deep learning in biomedicine. In: Nature biotechnology 36 (9), pp. 829–838. https://doi.org/10.1038/nbt.4233

Wilkinson, Mark et al. (2016): The FAIR guiding principles for scientific data management and stewardship. In: Scientific Data 3 (160018), pp. 1–9. https://doi.org/10.1038/sdata.2016.18

Zou, James; Huss, Mikael; Abid, Abubakar; Mohammadi, Pejman; Torkamani, Ali; Telenti, Amalio (2019): A primer on deep learning in genomics. In: Nature Genetics 51 (1), pp. 12–18. https://doi.org/10.1038/s41588-018-0295-5

REINHARD HEIL
is a philosopher and senior researcher at ITAS, KIT. His main research areas are social consequences of artificial intelligence and transhumanism/human enhancement.

DR. NILS B. HEYEN
is a sociologist and works as a senior researcher in the Competence Center Emerging Technologies at Fraunhofer ISI. Among his main research areas are technical and social innovations in medicine and health care as well as the science-society relationship.

MARTINA BAUMANN
is a member of the research group Life, Innovation, Health, and Technology at ITAS, KIT, where she focuses on ethical and social aspects of bio- and medical technologies.

DR. BÄRBEL HÜSING
is a biologist and works as a senior researcher in the Competence Center Emerging Technologies at Fraunhofer ISI. Her main research areas are innovations in biotechnology, bioeconomy and in the life sciences.

DR. DANIEL BACHLECHNER
is a social scientist and economist. He works as head of Advanced Data Analytics at Fraunhofer Austria. His research focuses on artificial intelligence methods, privacy-preserving technologies and the economics of data sharing and use.

PROF. DR. ULRICH SCHMOCH
is a senior researcher in the Competence Center Emerging Technologies at Fraunhofer ISI. His main research areas are patent and publication statistics and characteristics of emerging technologies.

DR. HARALD KÖNIG
is a senior researcher at ITAS, KIT, since 2011, after having worked as a research group leader in molecular biology. His work focuses on the assessment and governance of emerging biotechnologies.

RESEARCH ARTICLE

Mitarbeiterfreundliche Implementierung von KI-Systemen im Hinblick auf Akzeptanz und Vertrauen

Erarbeitung eines Forschungsmodells auf Basis einer qualitativen Analyse

Maria Jung, *Fachbereich Gesellschaftswissenschaften, Hochschule Darmstadt, Haardtring 100, 64295 Darmstadt, DE (maria.jung@h-da.de)*
Jörg von Garrel, *Fachbereich Gesellschaftswissenschaften, Hochschule Darmstadt, Darmstadt, DE (joerg.vongarrel@h-da.de)* *0000-0002-3617-1798*

Zusammenfassung • Der Einsatz künstlicher Intelligenz (KI) in produzierenden Unternehmen bietet Chancen und Potenziale. KI kann neben der Wettbewerbsfähigkeit von Betrieben auch die arbeitnehmerische Selbstbestimmtheit fördern. Derzeit wird die Bedeutung der Mitarbeiter für einen effektiven und effizienten Einsatz von KI-Systemen oftmals zu wenig berücksichtigt, da der Fokus überwiegend auf der Technologie liegt. Aus diesem Grund wurde eine qualitative Studie durchgeführt, die die beiden Faktoren „mitarbeiterfreundliche Implementierung" und „mitarbeiterfreundlicher Betrieb" in Bezug auf Akzeptanz und Vertrauen von KI-Systemen analysiert. Aus den Erkenntnissen wurde ein prozessorientiertes Forschungsmodell konzipiert, das auf der Adoptionstheorie von Rogers basiert und Einflüsse verschiedener Technologieakzeptanzmodelle sowie akzeptanz- und vertrauensfördernde Faktoren umfasst. Die Ergebnisse zeigen, dass eine wahrgenommene Arbeitserleichterung und sichtbare Erfolgserlebnisse zu einer erhöhten Handlungsakzeptanz beitragen können.

Employee-friendly implementation of AI systems in terms of acceptance and trust. *Development of a research model based on a qualitative analysis*

Abstract • *The use of artificial intelligence (AI) in manufacturing companies offers opportunities and potentials. In addition to the competitiveness of companies, AI can also promote self-determination by employees. Currently, too little attention is often paid to the importance of employees for effective and efficient use of AI systems as the focus is predominantly on the technology. Therefore, a qualitative study was conducted to analyze the two factors "employee-friendly implementation" and "employee-friendly operation" in terms of acceptance and trust of AI systems. From the findings, a process-oriented research model was developed based on Rogers' adoption theory and including influences from different technology acceptance models as well as factors promoting acceptance and trust. The results show that perceived ease of work and sense of achievement can contribute to increased acceptance of action.*

Keywords • *employee-friendly AI, AI acceptance, AI trust process*

Einleitung

Im Wandel hin zur Industrie 4.0 spielt der Einsatz von künstlicher Intelligenz (KI) eine zentrale Rolle, denn dieser bietet vielfältige Potenziale beziehungsweise Chancen für Unternehmen hinsichtlich der Steigerung ihrer Wettbewerbsfähigkeit (Stowasser et al. 2020, S. 5). Durch den Einsatz von KI-Systemen oder KI-basierten Prozessen in zahlreichen Unternehmensbereichen und Sektoren können Produktionsaktivitäten aber auch Geschäftsprozesse effektiv und effizient gestaltet werden (Rammer et al. 2020, S. 15). Häufig wird die Implementierung von KI-Systemen auf organisatorisch-technologischer Ebene betrachtet, die Perspektive der Arbeitnehmer[1] wird hierbei vernachlässigt. Infolgedessen muss insbesondere dem Vertrauensaufbau in Bezug auf Einführung und Nutzung innovativer Technologien eine besonders hohe Relevanz zugeschrieben werden, weil „[diese] die Balance der Kräfte zugunsten der Innovation umschlagen lassen oder diese daran hinder[t], Fahrt aufzunehmen" (Diekhöner 2018, S. 12).

1 Begriffe mit spezifischem Genus gelten im Sinne der Gleichbehandlung grundsätzlich für alle Geschlechter. Die verkürzte Sprachform beinhaltet keine Wertung.

https://doi.org/10.14512/tatup.30.3.37
Received: Jun. 14, 2021; revised version accepted: Oct. 18, 2021; published online: Dec. 20, 2021 (peer review)

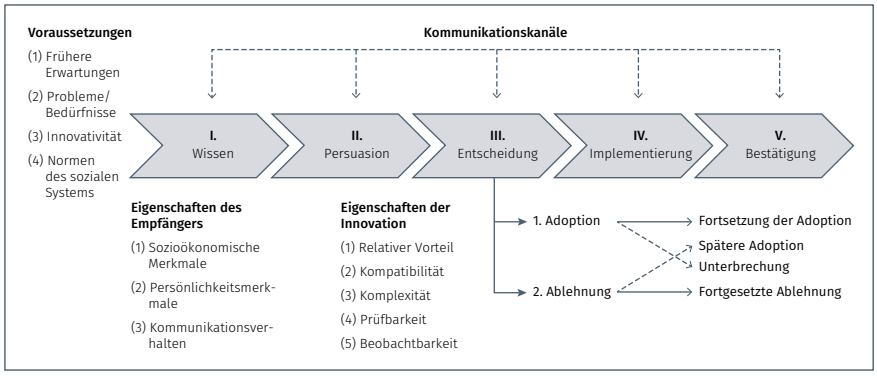


Abb. 1: Innovations-Entscheidungsprozess. *Quelle: nach Rogers 1983, S. 165*

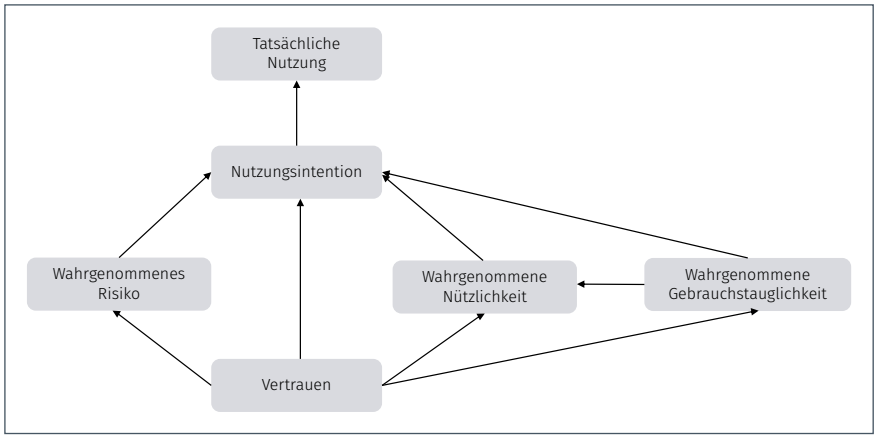


Abb. 2: Vereinfachtes Forschungsmodell zur Integration von Risiko, Vertrauen und TAM. *Quelle: nach Pavlou 2003, S. 122*

Ziel dieses Artikels ist daher die Erarbeitung eines Forschungsmodelles für eine mitarbeiterfreundliche Implementierung – hinsichtlich Vertrauens und Akzeptanz bei KI-Systemen in produzierende Unternehmen. Die Entwicklung dieses Modells erfolgt auf Basis der Erkenntnisse einer qualitativen Analyse unter Berücksichtigung der *Innovation Diffusion Theory* (IDT) (Rogers 1983), weiterer Akzeptanzmodelle (Pavlou 2003; Gefen et al. 2003; Backhaus 2017) sowie der Konzepte ‚Arbeitsfähigkeit' (Ilmarinen und Tempel 2002) und ‚Nutzerleben' beziehungsweise ‚User Experience' (Norman et al. 1995).

Theoretischer Rahmen

Um den Prozess der Akzeptanz- und Vertrauensbildung im Rahmen der Implementierung von KI-Systemen aus Mitarbeiterperspektive zu analysieren, bietet sich die recht abstrakte IDT im Sinne eines Innovations-Entscheidungsprozess (Adoptionsprozess) an. Dieser Ansatz ermöglicht eine subjektorientierte Betrachtung des Adoptionsprozesses einer Innovation und eignet sich somit potenziell für eine mitarbeiterzentrierte und prozessorientierte Implementierung von KI-Systemen. Idealerweise verläuft dieser Prozess vom ersten Bewusstwerden über eine Innovation bis zur bestätigten Nutzung dieser Innovation durch ein Individuum. Ausgangspunkt des Adoptionsprozesses sind (1) die Erfahrungen, (2) die Probleme bzw. Bedürfnisse, (3) die Innovationsneigung einer Person unter Berücksichtigung (4) der Normen des sozialen Systems (Rogers 1983).

Der Adoptionsprozess gliedert sich dabei in fünf Phasen. In der *ersten Phase* des Adoptionsprozesses wird sich das Individuum der Innovation bewusst, es erlangt Kenntnis über die Innovation und Wissen über ihre Funktionsweise (ibid., S. 164). Den Verlauf dieser Phase beeinflussen (1) Sozioökonomische Eigenschaften (z. B. Bildung), (2) Persönlichkeitsmerkmale (z. B. Offenheit) sowie (3) das Kommunikationsverhalten (z. B. Mediennutzung des Individuums) (ibid., S. 165 ff.). In der *zweiten Phase* erfolgt die individuelle Meinungsbildung und im Ergebnis die individuelle Einstellung zur Innovation. Nach Rogers (ibid, S. 14 ff.) üben folgende fünf Faktoren einen entscheidenden Einfluss auf diese Einstellung und auf die Wahrscheinlichkeit der Adoption einer Innovation aus:

1. Relativer Vorteil – Vergleich mit Althergebrachtem oder anderen Innovationen auf dem gleichen Gebiet;
2. Kompatibilität – Grad der Einpassung in bestehende Wertestrukturen, als Anknüpfung an bisherige Erfahrungen und Bedürfnisse der potenziellen Übernahme;
3. Komplexität – Unfähigkeit bzw. Unsicherheit, mit der Komplexität umzugehen;
4. Erprobbarkeit – bei Möglichkeit zum vorherigen Testen ist die Übernahme der Innovation wahrscheinlicher und geht schneller vonstatten;
5. Kommunizierbarkeit/Beobachtbarkeit – Sichtbarkeit/Kommunizierbarkeit der Innovation für andere Personen.

Dieser Phase wird eine besonders hohe Relevanz zugeschrieben, da sie den Grundstein für den weiteren Verlauf hinsichtlich Adoption oder Ablehnung der Innovation legt. In der anschließenden *dritten Phase* erfolgen Adoption oder Ablehnung der Innovation. Da der Adoptionsprozess zeitlich dynamischer Natur ist, ist er im Falle einer Ablehnung nicht notwendigerweise von dauerhafter Natur und es kann bei einem erneuten Durchlaufen des Adoptionsprozesses zu späteren Adoptionen kommen. Im Falle einer Adoption schließt die *vierte Phase* der Nutzung bzw.

der Implementierung der Innovation an, die durch die Demonstration im konkreten Kontext gekennzeichnet ist. Hierbei spielen die gesammelten Erfahrungen mit dem Produkt eine zentrale Rolle. Denn fallen diese positiv aus, so kommt es zur Bestätigung und folglich zur Adoption. Die *fünfte Phase* der Bestätigung ist charakterisiert durch eine Informationssuche nach Faktoren, die eine mögliche Dissonanz reduzieren und somit die getroffene Entscheidung unterstützen (Rogers 1983, S. 184). Die Adoption wird nicht fortgeführt, wenn überlegenere Innovationen zur Verfügung stehen oder das Individuum mit der Innovation unzufrieden ist (Rogers 1983, S. 209).

Mit Bezug auf das dargestellte Ziel einer akzeptanz- und vertrauensförderlichen Implementierung kann der Innovations-Entscheidungsprozess nach Rogers insgesamt als wichtiger Orientierungsrahmen zur Gestaltung der Einstellungs-, Handlungs- und Nutzungsakzeptanz gelten (Kollmann 1998). Um zusätzlich die Dimension der Gestaltung des Vertrauens im Mensch-Technik-Kontext zu berücksichtigen, bieten sich Modelle zur Technologieakzeptanz an (Backhaus 2017, S. 23).

Stehen im von Davis (1989) erarbeiteten Technologieakzeptanzmodell (TAM) die wahrgenommene Gebrauchstauglichkeit (*Perceived Ease of Use*) und die wahrgenommene Nützlichkeit des Systems (*Perceived Usefulness*) im Mittelpunkt, wurden in späteren Modellen – u. a. TAM 2 (Venkatesh und Davis 2000) oder TAM3 (Venkatesh und Bala 2008) – zusätzliche Einflussfaktoren (z. B. Geschlecht, Alter, Erfahrung mit dem System) ergänzt. Insbesondere das Modell nach Pavlou (2003) stellt in diesem Kontext einen häufig verwendeten Modellansatz dar, der das Konstrukt Vertrauen als Einflussvariable der Technologieakzeptanz berücksichtigt (siehe Abb. 2):

Strukturell wirken in diesem Modell – analog zum Technologieakzeptanzmodell – der wahrgenommene Nutzen sowie die wahrgenommene Gebrauchstauglichkeit positiv auf die Nutzungsintention. Pavlou (2003, S. 106 ff.) ergänzt in seinem Modell aber als weiteres, negativ zur Nutzungsintention korrelierendes Konstrukt das wahrgenommene Risiko. Alle diese vier Faktoren werden wiederum direkt vom (Nutzer-)Vertrauen beeinflusst: Vertrauen wirkt auf (1) den wahrgenommenen Nutzen, da erst durch das Vorhandensein eines (Nutzer-)Vertrauens eine effektive und effiziente Nutzung ermöglicht wird, (2) die wahrgenommene Gebrauchstauglichkeit, da durch Vertrauen die Komplexität der Situation reduziert werden kann, (3) das wahrgenommene Risiko, da vorhandenes (Nutzer-)Vertrauen risikomindernd wirkt sowie (4) die Nutzungsintention, da vorhandenes (Nutzer-)Vertrauen die Absicht der Nutzung erhöht. Welche Faktoren Einfluss auf das Vertrauen haben, verdeutlicht dieses Modell aber nicht.

Hinsichtlich der logischen Wirkbeziehungen zwischen (Nutzer-)Vertrauen und Technikakzeptanz besteht in der Literatur wenig Konsens. Anders als das genannte Modell von Pavlou kommen bspw. Gefen et al. (2003, S. 74 ff.) zu dem Schluss (siehe Abb. 3), dass sich die wahrgenommene Gebrauchstauglichkeit sowohl auf den wahrgenommenen Nutzen als auch auf das Vertrauen auswirkt, wobei das Vertrauen wiederum auch

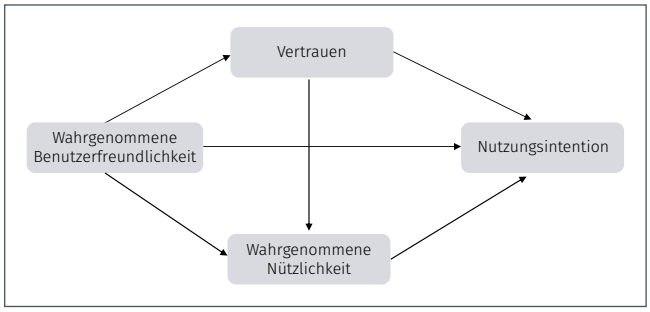


Abb. 3: Forschungsmodell zu Integration von TAM und Vertrauen.
Quelle: nach Gefen et al. 2003, S. 71

auf den wahrgenommenen Nutzen wirkt. Alle drei Faktoren beeinflussen dabei die Nutzungsintention. Gerade die wahrgenommene Gebrauchstauglichkeit beeinflusst damit das Vertrauen.

In dem Modell von Backhaus (2017) wird der Einfluss des Faktors Vertrauen auf die Technologieakzeptanz um das Konstrukt des Nutzerlebens erweitert. Im Gegensatz zur (wahrgenommenen) Gebrauchstauglichkeit erweitert Nutzererleben diese funktionale Perspektive um Erfahrungen und Erlebnisse eines Nutzers mit der Technologie. Backhaus kommt zu dem Ergebnis, dass Vertrauen stark mit dem Nutzererleben zusammenhängt und sich diese beiden Faktoren gegenseitig bedingen. Weiterhin wird Vertrauen durch die Eigenschaften des Technologieanbieters sowie der Technologie beeinflusst (Backhaus 2017, S. 187): „Zukünftige Studien sollten Vertrauen mit in die Untersuchungen einbeziehen und das Nutzererleben weiter fassen, so z. B. im Kontext des Kundenerlebens“ (Backhaus 2017, S. 88).

Im industriellen Sektor werden mit KI-Systemen meist automatisierte Entscheidungsfindungssysteme bezeichnet, welche den Menschen anwendungsorientiert unterstützen und auch als industrielle KI betitelt werden (Bundesregierung 2018, S. 22). Da die Art der Unterstützung zwischen Menschen und KI-Systemen vielfältiger Natur sein kann, muss – vor dem Hintergrund der genannten akzeptanz- und vertrauensgestaltenden Faktoren – eine detailliertere Auseinandersetzung mit unterschiedlichen User-Interfaces stattfinden (nach Bergstein 2017 in Apt und Priesack 2019, S. 232).

Ein hierfür geeigneter Ansatzpunkt bzw. Lösungsansatz kann das Konzept der Autonomiestufen in der industriellen Produktion bilden (BMWi 2019, S. 13 f.): Diese Autonomiestufen stellen die „[…] kontinuierliche Veränderung der Verantwortung des Anlagenbetriebs vom Menschen hin zum autonomen System“ (BMWi 2019, S. 13) dar. Der KI-bedingte Gesamtautonomiegrad eines industriellen Vorgangs kann sich von nicht-autonom (Stufe 0), über teil-autonom (Stufen 1–4), bis hin zu vollautonom (Stufe 5) unterscheiden. So ist in den ersten Stufen (Stufe 0–2) eine menschliche Einwirkung weiterhin erforderlich, so dass der Mensch nach wie vor die Kontrolle und die Verantwortung über das technologische System hat. Ab der dritten Stufe wächst der Anspruch an Zuverlässigkeit sowie Verantwortung des KI-Systems. Bei der höchsten Autonomiestufe (Stufe 5)

kann der Bediener komplett abwesend sein, weil keine Interaktion für den Betrieb der Anlage mehr erforderlich ist (BMWi 2019, S. 14 ff.). Trotz vieler Vorteile von KI weist diese Technologie dennoch Defizite auf. Dazu gehören beispielsweise die Vorkonfiguration von Daten, Umgebungen und Informationen, wodurch deutlich wird, dass die Auswirkungen von KI auf Beschäftigte und deren Arbeit weiterhin vom technologischen Fortschritt abhängen (nach Dengler und Matthes 2015 in Apt und Priesack 2019, S. 223).

Methodisches Vorgehen

Um ein Forschungsmodell für einen ‚KI-Vertrauensprozess' zur mitarbeiterfreundlichen Implementierung aufzubauen, ist es sinnvoll, neben (1) dem Adoptionsprozess nach Rogers, der den Prozess des Vertrauens- und Akzeptanzaufbaus – vom ersten Kennenlernen, über die Nutzung bis hin zu letztendlich regelmäßig bestätigten Handeln aus Subjektsicht darstellt, weiterhin (2) diverse Technologieakzeptanzmodelle mit einem Fokus auf Vertrauen (Pavlou 2003; Gefen et al. 2003; Backhaus 2017) sowie (3) das Konzept der Autonomiestufen in der industriellen Produktion heranzuziehen.

Auf dieser theoretischen Basis haben wir ein qualitatives Erhebungsinstrument erarbeitet, das verschiedene Szenarien berücksichtigt, um einer differenzierten Ausgestaltung hinsichtlich des Aufbaus und der Gestalt von Vertrauen und Akzeptanz mit Bezug auf verschiedene industrielle KI-basierte Arbeitssysteme gerecht zu werden. Das erste, moderate Szenario hat ein Arbeitssystem dargestellt, in dem eine Unterstützung von Mitarbeitern durch Objekterkennung und Ausgabe von Hinweisen auf einem Bildschirm erfolgt. Das zweite Szenario hat ein auf KI aufbauendes fahrerloses Transportsystem gezeigt, das eigenständig Tätigkeiten in einem kollaborativen Arbeitsprozess übernimmt. Szenario 3 hat als radikales Szenario autonom agierende KI-basierte Robotik-Lösungen im Bereich der Logistik vorgeführt.

Insgesamt wurden 15 Personen im Rahmen von Einzel- und Gruppeninterviews befragt, bei denen es sich um (1) Geschäftsführer sowie Manager, (2) KI-Experten sowie (3) Mitarbeiter aus der industriellen Produktion/Logistik handelte.[2] Pro Akteursgruppe wurden fünf Personen interviewt, siehe Abb. 4.

2 Die Interviews sind im April 2021 online durchgeführt und mit einem Aufnahmegerät aufgezeichnet worden. Anschließend ist die Transkription sowie die Auswertung mit dem Tool f4 erfolgt.

Ergebnisse

Die Auswertung der Ergebnisse ist induktiv erfolgt, um ein verifiziertes Forschungsmodell aufzubauen. Die Ergebnisse der qualitativen Studie legen nahe, dass eine erfolgreiche Implementierung von KI-Systemen in Unternehmen, unabhängig von der Autonomiestufe, den Phasen des Adoptionsprozesses nach Rogers folgt. So bildet den Ausgangspunkt die konkrete, unternehmensspezifische Problemlage in Kombination mit den Erfahrungen, die eine Organisation mit bestimmten Technologien gesammelt hat und in Verbindung mit finanziellen Restriktionen sowie technischen Determinanten (u. a. *Fit to System*, *Fit to Data*). Weitere adoptionsförderliche Vorbedingungen bilden gerade vor dem Hintergrund hoch-autonomer KI-Systeme die Innovationsfähigkeit der Organisation, eine innovationsfreundliche Unternehmenskultur sowie ein entsprechendes Engagement des Managements. Gerade die drei letztgenannten Faktoren können sich positiv auf das Vertrauen der Mitarbeiter auswirken, die das Unternehmen oder das Management weitestgehend als Vorbildfunktion anerkennen.

Um neue KI-Systeme erfolgreich einzuführen, betonen die befragten KI-Experten, Manager und Mitarbeiter die Notwendigkeit einer frühzeitigen Aufklärung seitens des Unternehmens. Schon in dieser frühen Phase sollten alle relevanten Organisationsmitglieder Kenntnis über das KI-System und Wissen über dessen Funktionsweise erlangen. Gerade Führungskräfte sollten hier proaktiv auf die Mitarbeiter zugehen und die geplante Implementierung von KI-Systemen thematisieren. In diesem Kontext sollten auch relevante Auswirkungen auf die Arbeitsfähigkeit der Mitarbeiter erörtert werden. Insbesondere die Darstellung der Tätigkeitsfelder der KI-Systeme und hieraus potenziell resultierender Tätigkeitsverluste bis hin zu Ausführungen zur Arbeitsplatzsicherheit sollten direkt thematisiert werden, um Existenzängste und eine damit verbundene Ablehnung der Neuerung zu vermeiden. Vor allem Mitarbeiter haben in den Interviews betont, dass ihnen die transparente Kommunikation der Arbeitsplatzsicherheit ein zentrales Anliegen ist. Um den Verlauf dieser Phase positiv zu beeinflussen, ist es auch sinnvoll, individuelle Eigenschaften der Mitarbeiter, u. a. sozioökonomische Eigenschaften oder auch Persönlichkeitsmerkmale wie Offenheit, sowie das Kommunikationsverhalten der Mitarbeiter als Gelingensbedingungen zu berücksichtigen und in eine zielgruppengerechte Kommunikation einfließen zu lassen. Dadurch kann schon vor dem Einsatz von KI die Basis für ein stärker individualisiertes Nutzererleben gelegt werden. Gerade transparente Kommunikation

Vertrauen in die Organisation als Vorbedingung ermöglicht ein positives Nutzererleben, was wiederum zu Vertrauen in KI-basierte Arbeitssysteme führt.

Management/ Führungsposition (M)	**Manager 1**	**Manager 2**	**Manager 3**	**Manager 4**	**Manager 5**
	Produktionsleiter bei einem Hersteller von Drähten und Kabelsystemen für Standard- und Spezialleitungen	**Geschäftsführer und Gesellschafter** bei einem Hersteller für Materialprüfmaschinen und Prüfsystemen	**Montageleiter** bei einem Hersteller für Materialprüfmaschinen und Prüfsystemen	**Manager** bei einem Hersteller für Materialprüfmaschinen und Prüfsystemen	**Manager (Strategie)** bei einem Anbieter von Produkten und Dienstleistungen in der medizinischen Versorgung
	(ca. 650 Mitarbeiter in Deutschland)	(ca. 80 Mitarbeiter in Deutschland)	(ca. 80 Mitarbeiter in Deutschland)	(ca. 80 Mitarbeiter in Deutschland)	(ca. 125.000 Mitarbeiter in über 20 Ländern)
KI-Experten (E)	**KI-Experte 1**	**KI-Experte 2**	**KI-Experte 3**	**KI-Experte 4**	**KI-Experte 5**
	Experte aus dem E-Commerce bei einem Süßwarenkonzern	**Geschäftsführer** bei einer Beratungsgesellschaft im Bereich Digitalisierung	**Consultingleiter** bei einem führendem Hard- und Softwareentwickler	**Projektleiter** bei einem deutschen Forschungsinstitut (Bereich KI)	**Projektleiter** bei einem deutschen Forschungsinstitut (Bereich KI)
	(ca. 7.000 Mitarbeiter in 14 Ländern)	(ca. 33 Mitarbeiter in Deutschland)	(ca. 150.000 Mitarbeiter in über 100 Ländern)	(ca. 200 Mitarbeiter in Deutschland)	(ca. 200 Mitarbeiter in Deutschland)
Mitarbeiter (A)	**Arbeiter 1**	**Arbeiter 2**	**Arbeiter 3**	**Arbeiter 4**	**Arbeiter 5**
	Lagerist bei einem Hersteller für Materialprüfmaschinen und Prüfsystemen	**Monteur** bei einem Hersteller für Materialprüfmaschinen und Prüfsystemen	**Produktionsmitarbeiter Fräsetechnik** bei einem Prüfmittelhersteller	**Produktionsmitarbeiter** bei einem Prüfmittelhersteller	**Abteilungsleiter** bei einem Prüfmittelhersteller
	(80 Mitarbeiter in Deutschland)	(80 Mitarbeiter in Deutschland)	(ca. 40 Mitarbeiter in Deutschland)	(ca. 40 Mitarbeiter in Deutschland)	(ca. 40 Mitarbeiter in Deutschland)

Abb. 4: Charakteristika der Interviewpartner. *Quelle: eigene Darstellung*

und Offenheit gegenüber Mitarbeitern durch das Management (als interpersonelle Faktoren) als auch Unternehmenskultur und Innovationsfähigkeit (als strukturelle Faktoren) sind als zentrale Erfolgsfaktoren genannt worden, die „zuversichtliche positive Erwartungen" (Oswald 2010, S. 63) gegenüber dem Einsatz KI-basierter Systeme im Sinne eines organisationalen Vertrauens aufbauen bzw. verstetigen sollen. Um die Meinungsbildung des Individuums zur KI und damit die Einstellungsakzeptanz positiv zu gestalten, sollte die Darstellung von Nützlichkeit des KI-Systems – im Sinne des Nettonutzens als Differenz zwischen Nutzen und Kosten bzw. Risiken – für jeden einzelnen Mitarbeiter im Fokus stehen. Die weitere Ausgestaltung des Nutzererlebens sollte Eigenschaften, Fähigkeiten und Funktionsweise der KI auf Arbeitsplatzebene beinhalten, aber auch deren Grenzen u. a. mit Bezug auf die jeweilige Autonomiestufe. Zentrale Faktoren aus Sicht der Interviewten sind in diesem Kontext, dass Faktoren wie Gestalt und Haptik des KI-Systems für die Mitarbeiter angenehm und intuitiv gestaltet sind und somit eine hohe, funktionale Gebrauchstauglichkeit gewährleistet ist. Gerade mit Bezug auf die unterschiedlichen Autonomiestufen in der industriellen Produktion bei KI-Systemen hat sich herausgestellt, dass das User Interface eines KI-Systems einen entscheidenden Einfluss auf die Einstellungs- und Handlungsakzeptanz haben kann. Durch eine klare Darlegung der Faktoren ist es für den Mitarbeiter möglich, sich eine individuelle Meinung zur Gebrauchstauglichkeit, aber auch zu Nutzen und Risiken des KI-Systems zu bilden. Neben den schon dargestellten Erfolgsfaktoren, die Vorteile der KI-Systeme klar darzulegen und diese kompatibel zu den bisherigen Arbeitstätigkeiten einfach anwendbar zu gestalten, erhöhen in dieser Phase insbesondere eine mögliche Erprobbarkeit und Beobachtbarkeit des KI-Systems die Wahrscheinlichkeit einer Adoption. Daher können Workshops aber auch Demonstratoren sinnvolle Elemente einer positiven Gestaltung des Nutzerlebens sein.

Eine positive Bewertung und somit auch eine positive Einstellungsakzeptanz führt zu einer individuellen Entscheidung durch den Mitarbeiter, das KI-System zu nutzen. Diese Handlungsakzeptanz kann dann zu einer Nutzungsakzeptanz führen, wenn die Erwartungen hinsichtlich Gebrauchstauglichkeit, Nutzen aber auch Risiken (Soll) mit der tatsächlich wahrgenommenen Situation (Ist) übereinstimmen, positiver bewertet werden und somit das Nutzererleben als überwiegend positiv bewertet wird.

Die Ergebnisse der Interviews verdeutlichen, dass vor allem die erfahrene Arbeitserleichterung sowie sichtbare Erfolgserlebnisse zu einer erhöhten Handlungsakzeptanz beitragen können. Gleichzeitig könnten Mitarbeiter durch eine mögliche Arbeitsentlastung profitieren, indem sie nun ‚sinnstiftende' Aufgaben erledigen. Werden die Erwartungen an das KI-System in der

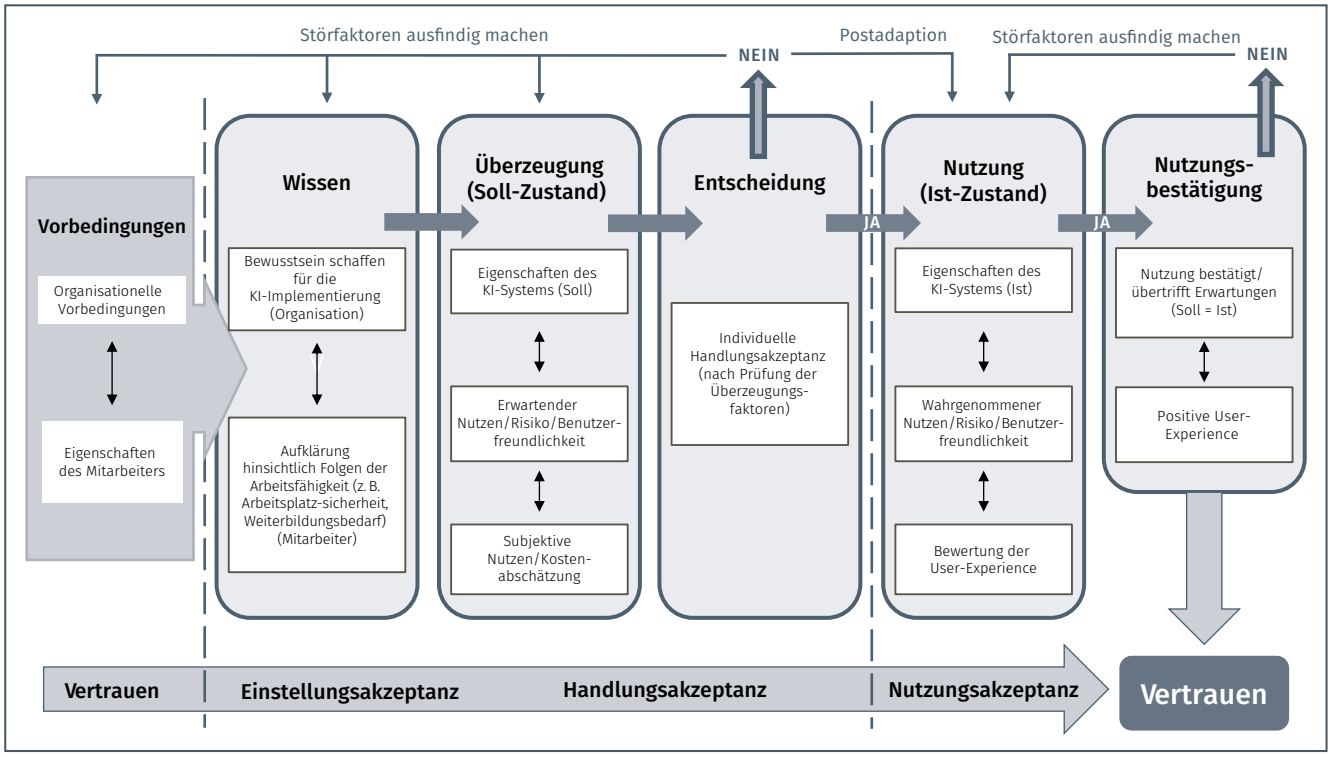


Abb. 5: Prozessorientiertes Forschungsmodell zur Schaffung von Akzeptanz und Vertrauen bei Mitarbeitern hinsichtlich KI. *Quelle: eigene Darstellung*

kontinuierlichen Nutzung erfüllt oder sogar übertroffen, führt dies zu einer Nutzungsbestätigung, einer Bestätigung des positiven Nutzerlebens und final zu einem Vertrauen in das KI-System. Unternehmen müssen in diesem Kontext darauf achten, Defizite am System oder sonstige Hindernisse schnellstmöglich zu eliminieren, da das gewonnene Vertrauen der Mitarbeiter durch Systemfehler sonst schnell wieder zunichtegemacht werden kann. In Summe führen die Ergebnisse damit zu einem prozessorientierten Forschungsmodell wie in Abb. 5 dargestellt.

Fazit und Ausblick

Um einen innovativen Vertrauensaufbau bei Mitarbeitern hinsichtlich neu einzuführender KI-Technologien zu generieren, kann neben diversen Technologie-Akzeptanzmodellen insbesondere der Adoptionsprozess nach Rogers (1983) hilfreich sein, um den Prozess des Vertrauens- und Akzeptanzaufbaus – vom ersten Kennenlernen, über die Nutzung, bis hin zum letztendlich regelmäßig bestätigten Handeln – zu gestalten und einen ‚KI-Vertrauensprozesses' im Sinne einer mitarbeiterfreundlichen Implementierung zu ermöglichen. Im Gegensatz zu den in diesem Artikel dargestellten Modellen (Pavlou, 2003; Gefen et al. 2003; Backhaus, 2017) scheint Vertrauen aber einerseits als organisatorisches Vertrauen (im Sinne eines interpersonellen und strukturellen Vertrauens) eine Vorbedingung zu sein, um ein positives Nutzerleben im Kontext der industriellen Produktion zu ermöglichen. Andererseits ist Vertrauen in das KI-System als zentrales Ergebnis dieses Prozesses anzusehen.

In einem nächsten Forschungsvorhaben wird auf Basis dieser qualitativ generierten Ergebnisse eine für die Praxis relevante Checkliste erarbeitet, die Hinweise für die betrieblichen Entscheider und Umsetzer liefern soll, um KI-Systeme erfolgreich im Unternehmen zu implementieren. Weiterhin werden quantitative Studien zur detaillierteren Analyse der Faktoren Vertrauen und Akzeptanz im Rahmen der Implementierung und Nutzung von KI-Systemen durchgeführt.

Angabe von Finanzierungsquellen

Das diesem Bericht zugrundeliegende Vorhaben wurde mit Mitteln des Bundesministeriums für Bildung und Forschung unter dem Förderkennzeichen 02L19C157 gefördert. Die Verantwortung für den Inhalt dieser Veröffentlichung liegt bei den Autoren.

Literatur

Apt, Wenke; Priesack, Kai (2019): KI und Arbeit. Chance und Risiko zugleich. In: Volker Wittpahl (Hg.): Künstliche Intelligenz. Technologie, Anwendung, Gesellschaft. Berlin: Springer Vieweg, S. 221–238. https://doi.org/10.1007/978-3-662-58042-4_14

Backhaus, Nils (2017): Nutzervertrauen und -erleben im Kontext technischer Systeme. Empirische Untersuchungen am Beispiel von Webseiten und Cloudspeicherdiensten. Dissertation, Technische Universität Berlin. Online verfügbar unter https://d-nb.info/1156183804/34, zuletzt geprüft am 13. 10. 2021.

BMWi – Bundesministerium für Wirtschaft und Energie (Hg.) (2019): Technologieszenario „Künstliche Intelligenz in der Industrie 4.0". Working Paper. Berlin: Plattform Industrie 4.0. Online verfügbar unter https://www.plattform-i40.de/IP/Redaktion/DE/Downloads/Publikation/KI-industrie-40.pdf?__blob=publicationFile&v=10, zuletzt geprüft am 13. 10. 2021.

Bundesregierung (2018): Strategie Künstliche Intelligenz der Bundesregierung. O. O.: o. V. Online verfügbar unter https://www.bmwi.de/Redaktion/DE/Publikationen/Technologie/strategie-kuenstliche-intelligenz-der-bundesregierung.pdf?__blob=publicationFile&v=10, zuletzt geprüft am 13. 10. 2021.

Davis, Fred (1989): Perceived usefulness, perceived ease of use and user acceptance of information technology. In: MIS Quarterly 13 (3), S. 319–340. https://doi.org/10.2307/249008

Diekhöner, Philipp Kristian (2018): The trust economy. Wiesbaden: Springer. https://doi.org/10.1007/978-3-662-57459-1

Gefen, David; Karahanna, Elen; Straub, Detmar (2003): Trust and TAM in online shopping. An integrated model. In: MIS Quarterly 27 (1), S. 51–90. https://doi.org/10.2307/30036519

Ilmarinen, Juhani; Tempel, Jürgen (2002): Arbeitsfähigkeit 2010. Was können wir tun, damit Sie gesund bleiben? Hamburg: VSA.

Kollmann, Tobias (1998): Akzeptanz innovativer Nutzungsgüter und -systeme. Konsequenzen für die Einführung von Telekommunikations- und Multimediasystemen. Wiesbaden: Springer Gabler. https://doi.org/10.1007/978-3-663-09235-3

Norman, Donald; Miller, Jim; Henderson, Austin (1995): What you see, some of what's in the future, and how we go about doing it. HI at Apple computer. In: Conference Companion on Human Factors in Computing Systems. New York: Association for Computing Machinery, S. 155. https://doi.org/10.1145/223355.223477

Oswald, Margit (2011): Vertrauen in Organisationen. In: Martin Schweer (Hg.): Vertrauensforschung. State of the art. Frankfurt am Main: Peter Lang, S. 63–85.

Pavlou, Paul (2003): Consumer acceptance of electronic commerce. Integrating trust and risk with the technology acceptance model. In: International Journal of Electronic Commerce 7 (3), S. 101–134. https://doi.org/10.1080/10864415.2003.11044275

Rammer, Christian; Bertscheck, Irene; Schuck, Bettina; Demary, Vera; Goecke, Henry (2020): Einsatz von Künstlicher Intelligenz in der Deutschen Wirtschaft. Stand der KI-Nutzung im Jahr 2019. Berlin: Bundesministerium für Wirtschaft und Energie.

Rogers, Everett (1983): Diffusion of Innovations. New York, NY: The Free Press.

Stowasser, Sascha et al. (2020): Einführung von KI-Systemen in Unternehmen. Gestaltungsansätze für das Change-Management (Whitepaper). München: Lernende Systeme – Die Plattform für Künstliche Intelligenz.

Venkatesh, Viswanath; Bala, Hillol (2008): Technology acceptance model 3 and a research agenda on interventions. In: Decision Sciences 39 (2), S. 273–315. https://doi.org/10.1111/j.1540-5915.2008.00192.x

Venkatesh, Viswanath; Davis, Fred (2000): A theoretical extension of the technology acceptance model. Four longitudinal field studies. In: Management Science 46 (2), S. 186–204. https://doi.org/10.1287/mnsc.46.2.186.11926

MARIA JUNG
studierte Angewandte Sozialwissenschaften an der Hochschule Darmstadt und ist seit Januar 2021 wissenschaftliche Mitarbeiterin im BMBF-geförderten Projekt *Kompetenzzentrum für Arbeit und Künstliche Intelligenz (KompAKI)*.

PROF. DR. JÖRG VON GARREL
hat eine Professur für Prozess- und Produktinnovationen an der Hochschule Darmstadt inne. Seine Forschungen fokussieren auf eine partizipative sowie effektive und effiziente Gestaltung von Arbeitssystemen und Arbeitsprozessen vor dem Hintergrund aktueller Entwicklungen der Digitalisierung und des demografischen Wandels.

OPEN ACCESS oekom

RESEARCH ARTICLE

"Don't let me be misunderstood"

Critical AI literacy for the constructive use of AI technology

Stefan Strauß, *Institut für Technikfolgen-Abschätzung (ITA), Österreichische Akademie der Wissenschaften, Apostelgasse 23, 1030 Wien, AT (sstrauss@oeaw.ac.at)* *0000-0003-1877-2415*

Abstract • Research and development as well as societal debates on the risks of artificial intelligence (AI) often focus on crucial but impractical ethical issues or on technocratic approaches to managing societal and ethical risks with technology. To overcome this, more practical, problem-oriented analytical perspectives on the risks of AI are needed. This article proposes an approach that focuses on a meta-risk inherent in AI systems: deep automation bias. It is assumed that the mismatch between system behavior and user practice in specific application contexts due to AI-based automation is a key trigger for bias and other societal risks. The article presents the main factors of (deep) automation bias and outlines a framework providing indicators for the detection of deep automation bias ultimately triggered by such a mismatch. This approach intends to strengthen problem awareness and critical AI literacy and thereby create some practial use.

„Don't let me be misunderstood". *Kritische KI-Kompetenz für den konstruktiven Umgang mit KI-Technologie*

Zusammenfassung • *Gesellschaftlicher Diskurs sowie Forschung und Entwicklung zu Risiken künstlicher Intelligenz (KI) fokussieren oft einseitig entweder auf praxisferne ethische Aspekte oder auf technokratische Ansätze zur Bewältigung gesellschaftlicher Risiken allein durch Technologie. Es bedarf jedoch praktikabler, problemorientierter Perspektiven. Dieser Beitrag konzentriert sich daher auf ein zentrales Meta-Risiko von KI-Systemen: Deep Automation Bias. Es wird davon ausgegangen, dass Diskrepanzen zwischen Systemverhalten und Nutzungspraktiken in bestimmten Anwendungskontexten aufgrund KI-basierter Automatisierung zentrale Auslöser von Bias und gesellschaftlichen Risiken sind. Der Beitrag stellt zentrale Faktoren von (Deep) Automation Bias vor und entwickelt einen analytischen Rahmen mit Indikatoren zur Erkennung von Diskrepanzen in KI-Systemen. Dieser Ansatz will durch Stärkung von Problembewusstsein und kritischer KI-Kompetenz auch praktischen Nutzen erzielen.*

https://doi.org/10.14512/tatup.30.3.44
Received: Jun. 14, 2021; revised version accepted: Oct. 18, 2021; published online: Dec. 20, 2021 (peer review)

Keywords • *deep automation bias, AI assessment, machine learning, uncertainty, awareness*

Introduction

The hype around artificial intelligence (AI) is yet unbroken. Machine learning (ML) algorithms gain influence on economic, social and political decisions affecting individuals directly and indirectly. Accordingly, there is a scientific and political debate on how to tackle the various ethical risks of a broader use of AI. These discussions are, though, mostly dominated by either general ethical issues such as human versus machine autonomy, matters of trust, fairness, accountability and transparency (FAT) or on technical solutions to avoid algorithmic discrimination. Correspondingly, there is a number of guidelines for "ethical AI" or "trustworthy AI" issued by the EU Commission's high-level expert group on AI and others (Floridi et al. 2018; HLEG 2019; AlgorithmWatch 2019; Hallensleben 2020). And a growing community deals with developing technical solutions for de-biasing and FAT-ML, for example in the annual ACM-FAT conferences (Selbst et al. 2019; Wieringa 2020; Eid et al. 2021). Without doubt, this involves various relevant research and development activities.

But there is also a certain gap between important but impractical ethical concepts on the one side and technocratic approaches to fix societal problems with algorithms on the other. Not without irony, this situation could even reinforce the myriad of AI-related risks ranging from bias and discrimination, lacking transparency, erosion of privacy and security, loss of autonomy etc. There is thus need for a broader debate and problem-oriented approaches on how to effectively comprehend and conceptualize socio-technical risks related to AI.

A main argument of this paper[1] is that AI-based automation plays a particular role here. To explore the risks of AI thus requires a stronger analytical focus on automation. To facilitate

1 Parts of this paper represent a condensed and modified version of Strauß (2021).

this, I identified deep automation bias (DAB) as a meta-risk of the societal use of AI entailing further risks. DAB is a multidimensional, wicked problem inherent to AI technology alluding to progress in deep learning and self-optimizing algorithms (Strauß 2018, 2021). The aim is to develop this concept of DAB further and propose it as part of a problem-oriented assessment framework of AI. The premise here is that essentially, AI-based technology represents a socio-technical system that fosters automation at different levels. Bias can result from pre-existing prejudice during technical development, technical issues like poor data quality, insufficient models or inappropriate operation of ML-algorithms; but also from rule conflicts between AI design and AI application contexts due to complexity gaps between statistical assumptions in the system and user practices. In each case, the common denominator is automation, though, on different socio-technical levels.

To understand how these levels interact requires a multilayer view on the interplay between design and use of AI technology which together shape societal impacts. The main focus of the paper is thus on how to improve the analytical perspective on AI as a socio-technical issue to foster the basic understanding and awareness on the related societal challenges. This is a contribution towards what I call 'critical AI literacy' to avoid the fallacy of seeking for technological fixes for societal problems. The paper is structured as follows: after this introduction, section two briefly discusses why AI bears wicked problems which cannot be addressed with technical means only. Section three then sheds light on critical AI literacy and the role of automation. Based on main factors affecting DAB section four sketches a problem-oriented assessment framework. Section five presents a short summary and concluding remarks.

Wicked problems require more than fairness, accountability and transparency

As several scholars argue, there is need for alternative socio-technical approaches to better grasp the societal and ethical issues of AI (Edwards and Veale 2017; Selbst et al. 2019; Tsamados et al. 2020). This is particularly relevant as the use of AI systems can involve and reinforce so-called wicked problems (Strauß 2021). They are a "'class of social system problems which are ill-formulated, where the information is confusing, where there are many clients and decision makers with conflicting values, and where the ramifications in the whole system are thoroughly confusing'" (Churchman 1967 cited in Buchanan 1992, p. 15). Wicked problems bear tensions between the *artificial* and the *natural* (ibid.). This basic conflict can reinforce with the use of AI, particularly due to its high degree of automation: AI transforms decision-making and entails risks of reducing *natural* aspects of society to machine-readable data models that are interpretable by *artificial* algorithms.

Bias in ML is a wicked problem inherent to AI. However, unbiasing and fostering FAT is not sufficient to avoid the related risks of undetected failure, self-fulfilling prophecies and an incremental normalization of AI biases in society. Sheer techno-fixes could even intensify these risks. Research on FAT and bias in ML is dominated by debates on how different types occur, i. e., preexisting, technical or emergent bias and how to avoid that AI and algorithms lead to discrimination and injustice (Friedman and Nissenbaum 1996; Simon et al. 2020; Wieringa 2020). This is important work but there is a tendency to frame this socio-technical issue as a technological one or to get lost in general ethical debates on fairness, justice etc. and seeking technical solutions to ethical problems. This can be counterproductive. Unbiasing approaches, e. g., with adaptive algorithms, may increase complexity and opacity of AI which further reinforce societal risks.

Obviously, not just the technical design of AI is relevant but in particular, how (in-)compatible the technical system is with its socio-technical application contexts. This is crucial to tackle the risks of AI, which requires a broader, problem-oriented perspective that fosters analytical views on both, technical and societal issues of AI systems. To circumvent one-sided views, like ethical debates beyond practicability or misleading technocratic approaches and underestimation of risks, raising problem-awareness among decision-makers and persons interacting with AI systems is essential. However, as of yet, there is a lack of awareness and analytical perspectives in this regard. I thus suggest to focus more on the specific role of automation in AI and how to establish what I call here critical AI literacy. Critical AI literacy here means the ability to comprehend the core features of an AI system and its (in-)compatibility with its particular application contexts in a (necessarily) more complex sociotechnical reality.

Beyond practicability or misleading technocratic approaches and underestimation of risks, raising problem-awareness among decision-makers and persons interacting with AI systems is essential.

Critical AI literacy: understanding AI-based automation (bias)

A crucial question for the use of AI is whether it matches with the requirements of a particular application context. This implies that the contextual environment of an AI system affects the occurrence of bias. Tsamados et al. (2020) discuss context bias on the example of a healthcare system for resource management in hospitals. The system may function properly for one hospital that fits to the model the system uses but may cause problems in others, e. g., rural clinics with different contextual factors. But as argued, at the core, the various risks of AI ultimately derive from conflicts due to different forms of automation. Automation bias (AB) is the general risk of uncritically accepting the outcome of an automated system (Goddard et al. 2012, 2014). AI intensifies this risk and thus DAB represents a meta-risk of AI. The following examples highlight this:

Even very simple forms of automation can cause serious problems as the case of the automated renaming function in Excel tables shows: studies detected failure rates of 20 per cent implying that every third table containing genetic data presents false information as gene names are automatically renamed to dates (e. g., MARCH1 to 1-Mar). Abeysooriaya et al. (2021) show that this problem still exists and recommend human workarounds. Thus, even simple errors may create severe impact. Particularly, if these errors remain undetected and are processed further by AI systems.

Imagine an autopilot-system of an airplane, a classical form of automation. Basically, it is a rule-based system which functions with sensors and real-time data on geolocation, weather etc. Hence it needs a plausible data model of the plane's environment and reliable information on its behavior so that the human pilot can monitor if autopilot and plane operate as intended and can intervene immediately in case of problems. Any hidden error like a faulty label in a data table could threaten human lives. Recent cases of military drones autonomously attacking soldiers in Libya highlight that this is not a sheer theoretical risk (Hambling 2021). AB is a known risk of autopilots (Parasuraman et al. 2010; Goddard et al. 2014), mitigated with extensive training and technical features to improve controllability and avoid overreliance on the system. A precondition here is the basic predictability of system behavior and comprehensible rules determining its functionality. Hence system complexity must remain manageable. An autopilot that would permanently try to optimize a flight (e. g. with some predictability algorithm) without effective human intervention would be uncontrollable. The system would lever out human autonomy and agency and the conflict between system behavior and human intervention could escalate at any time. Tackling this risk requires more than transparency, accountability or explicability and is impossible without plausibility, reliability, predictability and effective intervenability to comprehend and correct the automated system.

Further examples are AI systems for job applications which evidently led to discrimination in various cases. As Harwell (2019) illustrates, the hiring platform HireVue calculated an "employability score" based on various data on job applicants including facial expressions and speech. Critics filed complaint and argued the system is biased, unfair and deceptive as it discriminates, e. g., due to different facial looks and spoken accents. Another system uses background images in applicants' portray photos to predict job qualification (Harlan and Schnuck 2021). For example, a person standing in front of a bookshelf then has higher chances to get a job offer for certain job sectors than a person submitting a photo with plain background. Obviously, skin color, ethnicity and background images have no relevance for a person's qualification. But people of color or persons with lower contrasting background images generally get lower scores. Hence the system reinforces racial and other forms of discrimination. This is an evident issue of various other AI systems, too. Various cases (O'Neil 2016; Borgesius 2018; Obermeyer et al. 2019; Köchling and Wehner 2020) demonstrate, how problematic it can be to automate social domains with AI. They underline the risk of DAB which is inevitable here if neither job applicants nor recruiters are unaware of the problem and no countermeasures to avoid discrimination are set. In any case, the system behaves unfair and unreliable.

Technical fixes like de-biasing to fix deficient image processing do not solve such problems as they are more than just technical issues. A typical technical solution to the above-mentioned bias would be to modify the algorithm so that it excludes image backgrounds when calculating a qualification score. This may ease bias resulting from images but any other problems with criteria the algorithm may process (e. g., ethnic facial features, residential district) would remain unsolved. Also, FAT is ineffective as transparency on the issue would not prevent from discrimination. Moreover, there are various cases of bias or stereotyping in data models and ML approaches with problematic effects on system behavior. Particularly sensitive is the use of AI in the health domain. Several studies reveal problems and unintended effects of decision-making systems here (Goddard et al. 2012; Cabitza et al. 2017; Gianfrancesco et al. 2018; Obermeyer et al. 2019). Gianfrancesco et al. (2018, p. 5) analyzed ML algorithms in clinical applications and found serious issues such as "overreliance on automation, algorithms based on biased data, and algorithms that do not provide information that is clinically meaningful". They conclude that easing these problems requires better understanding of AI and their ML approaches and corresponding measures to achieve this.

Towards a problem-oriented assessment framework

To raise problem-awareness, it is essential to understand how AI-based automation operates and how DAB occurs. The factors and framework presented here are meant as an approach to improve critical AI literacy. Basically, AI becomes problematic when there is a mismatch between system behavior and user

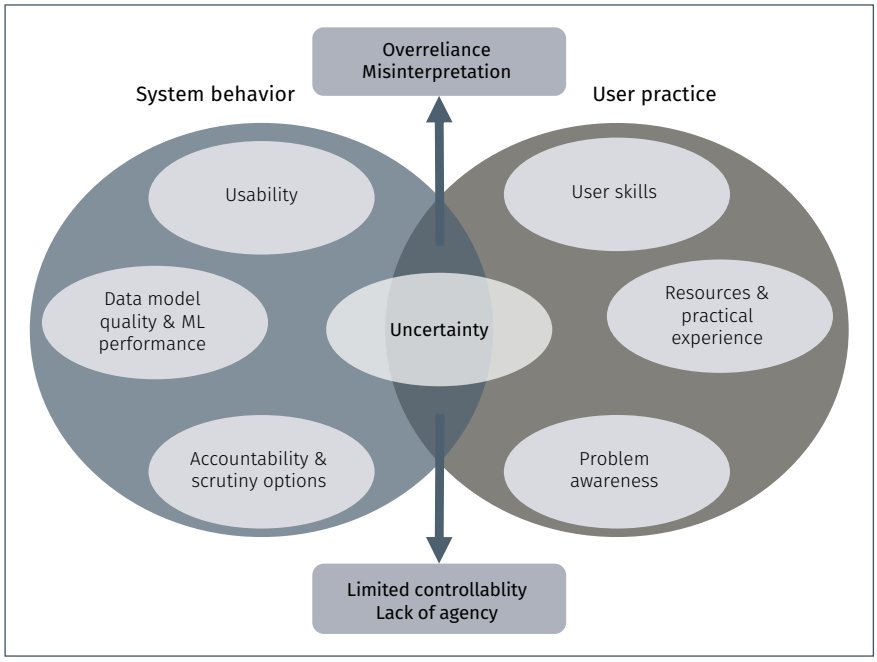


Fig. 1: Factors affecting DAB. *Source: Strauß 2021, p. 8*

practice in specific application contexts. To comprehend the meaning of mismatch is a precondition for the assessment of AI-related risks. This implies to understand the peculiarities of AI-based automation. Because irrespective of specific features, every AI system uses some form of automation and bears risks of (D)AB. But as Tsoukiàs (2020) reminds, automation is not inevitable with AI. It is a choice that needs to be legitimated and not an end in itself. It is crucial to scrutinize the automation approach of an AI system when assessing its impact. As a first step to sharpen the analytical lens, I suggest to conceptualize DAB as meta-risk of AI from which other risks derive from.

Main determinants of DAB

DAB addresses the problem of increasing complexity and opacity of AI technology that reinforce AB due to its dynamic, unpredictable and thus potentially uncontrollable behavior (Strauß 2018, 2021). Sophisticated ML approaches with features to self-optimization like deep learning or other forms of unsupervised learning aggravate this problem. It can extend the gap between human propensity to blindly trust AI technology and the limits of technology to match social complexity due to necessarily reductionistic models of society. The wider risk here is that dependence of society on AI systems reaches a stage where automated decisions – no matter if socially acceptable, ethical, correct or false – become uncontrollable. AI then constantly reshapes society and individual lives without effective alternatives. Figure 1 shows basic factors affecting DAB.

DAB bears at least two main risks: at the top is the "classical" risk of overreliance on the behavior of an AI system and misinterpretation; the bottom shows the additional risk of limited controllability and lack of agency. Both dimensions are interrelated and the severity of DAB depends on the interplay of different factors. The main connecting factor is uncertainty, shaped by technical as well as social issues that can reinforce mutually. System behavior strongly depends on the quality of data models and ML performance, usability, accountability and scrutiny options in the system. The social perspective involves user practices: various studies show that user skills, practical experience, resources (e. g. user knowledge to interpret a system, time and pressure to act), workload and effective options to scrutinize automated procedures affect AB (Goddard et al. 2012, 2014; Lyell and Coiera 2016). DAB further complicates these factors as AI increases system complexity, opacity and decreases options to scrutinize its functionality. Consequently, controllability, agency and options to intervene into automated decision-making can also decrease. The interplay of all these factors affect the severity of DAB and related risks. It is particularly higher, when the system lacks in options to scrutinize its behavior and/or problem awareness of the human user is low. For example, lacking accountability reduces agency and low problem awareness limits the user's ability to scrutinize, which again limits agency (Strauß 2021).

How to assess system behavior

Figure 2 sketches a three-level framework with indicators to identify DAB-related risks. The basic idea is to provide a simple checking tool to detect obscurities in system behavior. The focus is on the operational level as DAB risks become most apparent here.

The four main indicators (explicability, validity, plausibility and acceptability) represent a toolbox to check if the system operates properly. The related guiding questions apply to the whole operation process from input, output to action or decision. Any case of doubt or uncertainty triggers a more detailed review at all levels to uncover eventual mismatch between system behavior and user practice, technical flaws, or legal or ethical problems such as violation of human rights, discrimination etc. If all indicators point to normal system behavior, no DAB risk was detected. But regular detailed reviews of system behavior including all levels are advisable to avoid instability or any other issues. Obviously, all assessment levels are intertwined, but the schematic distinction supports comprehension of whether and how DAB occurs. If the system malfunctions because the data model is biased, then the system as a whole is biased. If the data model is OK but the system behavior is implausible there might be a

different reason; and if a system is basically explicable, its outcome is not necessarily ethical. For users interacting or testing AI, ethical reviews are impractical. It makes little sense to ask, for example, whether the system affects autonomy during operation. But it makes sense having some indicators to assess how exposed system behavior is to DAB-related risks. Acceptability is thus drawn at the intersection between operational and ethical level, because an unacceptable outcome during operation may indicate a severe legal or ethical issue which then needs to be analyzed further.

Briefly, applying the framework to the afore mentioned case of AI for applications underlines the necessity of different assessment levels. In a sheer technical sense, there might be no problem observable. Consider a typical ML framework with preexisting external data models embedded in the system from a trustful source. Without technical and operational assessment, revealing bias in the system, for example due to its mode of image processing, is impossible. To effectively assess how the system operates requires knowledge about technical features and how they affect the preselection of job applicants. The assessment would then show that the system is neither explicable nor valid, nor plausible, nor acceptable. Consequently, a deeper, ethical assessment would be triggered which then would reveal that the system approach in total is unethical as it discriminates persons based on irrelevant data. More detailed testing of the frameworks' practicability, as presented in this article, is subject to further research.

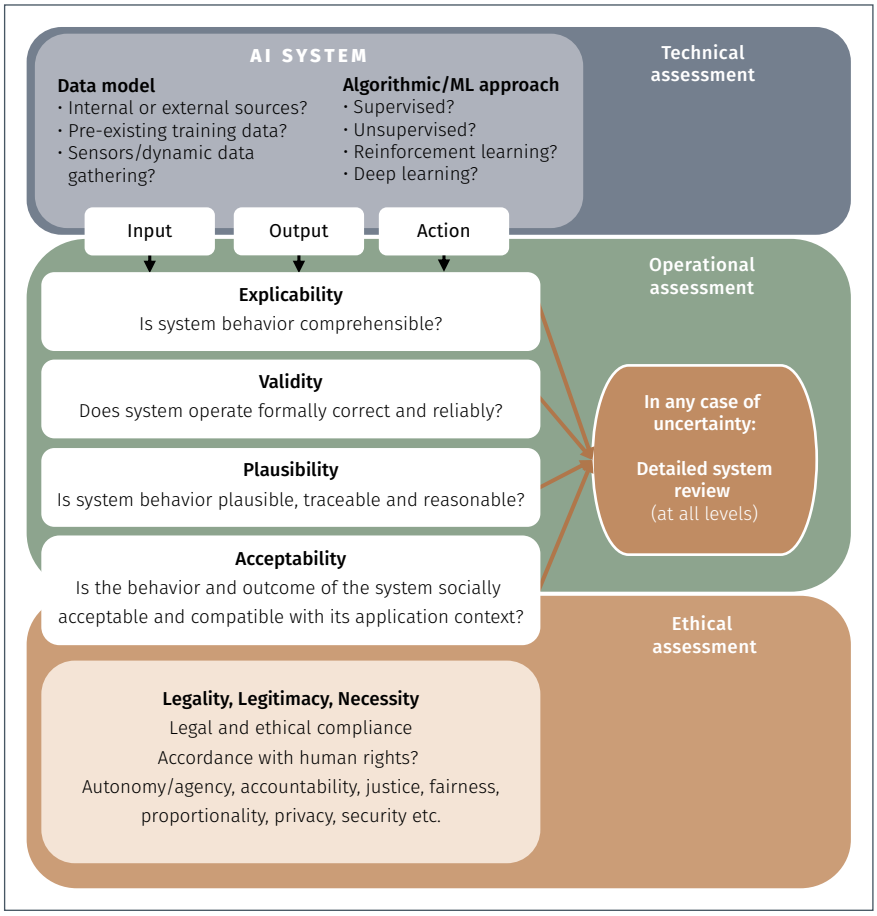


Fig. 2: A problem-oriented assessment framework. *Source: enhanced version of Strauß 2021, p. 9*

Concluding remarks

„I am just a soul whose intentions are good; oh lord please don't let me be misunderstood". This refrain of the famous song, first interpreted by Nina Simone in 1964, highlights the societal dilemma of the broader use of AI: (mis-)understanding is a key issue. AI bears high potential to transform society and there are many "good" intentions behind AI-based innovation. But good intentions, such as causing no harm and creating benefit, are not enough to tackle the myriad of risks and prevent AI from becoming a severe threat to society. The crux are misconceptions between technology design, specific application contexts and individual, institutional, societal and ethical requirements. Attempts to resolve them by integrating ethics into AI systems is like trying to explain to an automated vehicle why it should not cause accidents. This is doomed to fail because machines, irrespective of the degrees of their mechanical or digital automation processes, remain machines. Any such attempt aggravates ethical problems. To avoid that humans become "slave to the algorithm" (Edwards and Veale 2017, p. 1) we need more knowledge on the risks of AI and better strategies to cope with them. As a step in this direction, this paper suggests to foster critical AI literacy based on a problem-oriented approach with an explicit focus on DAB-related risks as trigger for further risks of AI. This approach is meant as awareness-raising tool which may also be of some practical use. The intention behind it is rather simple: to envision non-dystopian futures requires novel perspectives on AI to overcome technocratic approaches and revitalize humanistic perspectives on how to deal with AI in a constructive, socially acceptable manner. This can only work if all stakeholders, engineers, designers, policy makers, users and other persons concerned are aware of the factual risks and find ways to reduce them.

Acknowledgement

This research article has not received any external funding.

References

Abeysooriya, Mandhri; Soria, Megan; Kasu, Mary; Ziemann, Mark (2021): Gene name errors. Lessons not learned. In: PLoS Computational Biology 17 (7), p. e1008984. https://doi.org/10.1371/journal.pcbi.1008984

AlgorithmWatch (2019): Automating society. Taking stock of automated decision-making in the EU. Available online at https://www.algorithmwatch.org/automating-society, last accessed on 12. 10. 2021.

Borgesius, Frederik (2018): Discrimination, artificial intelligence, and algorithmic decision-making. Study for the Council of Europe. Strasbourg: DG of Democracy.

Buchanan, Richard (1992): Wicked problems in design thinking. In: Design Issues 8 (2), pp. 5–21. https://doi.org/10.2307/1511637

Cabitza, Federico; Rasoini, Raffaele; Gensini, Gian (2017): Unintended consequences of machine learning in medicine. In: JAMA 318 (6), pp. 517–518. https://doi.org/10.1001/jama.2017.7797

Edwards, Lilian; Veale, Michael (2017): Slave to the algorithm? Why a 'right to explanation' is probably not the remedy you are looking for. In: Duke Law & Technology Review 16 (1), pp. 18–84. https://doi.org/10.2139/ssrn.2972855

Eid, Fatma-Elzahraa et al. (2021): Systematic auditing is essential to debiasing machine learning in biology. In: Communications Biology 4 (183), p. 1–9. https://doi.org/10.1038/s42003-021-01674-5

Floridi, Luciano et al. (2018): AI4People – an ethical framework for a good AI society. Opportunities, risks, principles, and recommendations. In: Minds & Machines 28, pp. 689–707. https://doi.org/10.1007/s11023-018-9482-5

Friedman, Bataya; Nissenbaum, Helen (1996): Bias in computer systems. In: ACM Transactions on Information Systems, 14 (3), pp. 330–347. https://doi.org/10.1145/230538.230561

Gianfrancesco, Milena et al. (2018): Potential biases in machine learning algorithms using electronic health record data. In: JAMA Internal Medicine 178 (11), pp. 1544–1547. https://doi.org/10.1001/jamainternmed.2018.3763

Goddard, Kate; Roudsari, Abdul; Wyatt, Jeremy (2012): Automation bias. A systematic review of frequency, effect mediators, and mitigators. In: Journal of the American Medical Informatics Association 19 (1), pp. 121–127. https://doi.org/10.1136/amiajnl-2011-000089

Goddard, Kate; Roudsari, Abdul; Wyatt, Jeremy (2014): Automation bias. Empirical results assessing influencing factors. In: International Journal of Medical Informatics 83 (5), pp. 368–375. https://doi.org/10.1016/j.ijmedinf.2014.01.001

Hallensleben, Sebastian et al. (2020): From principles to practice. An interdisciplinary framework to operationalise AI ethics. Gütersloh: Bertelsmann Stiftung. Available online at https://www.bertelsmann-stiftung.de/fileadmin/files/BSt/Publikationen/GrauePublikationen/WKIO_2020_final.pdf, last accessed on 12. 10. 2021.

Hambling, David (2021): Drones may have attacked humans fully autonomously for the first time. In: New Scientist, 27. 05. 2021. Available online at https://www.newscientist.com/article/2278852-drones-may-have-attacked-humans-fully-autonomously-for-the-first-time/, last accessed on 12. 10. 2021.

Harlan, Elisa; Schnuck, Oliver (2021): Objective of biased? On the questionable use of artificial intelligence for job applications. Available online at https://web.br.de/interaktiv/ki-bewerbung/en/, last accessed on 12. 10. 2021.

Harwell, Drew (2019): A face-scanning algorithm increasingly decides whether you deserve the job. In: Washington Post, 06. 11. 2019. Available online at www.washingtonpost.com/technology/2019/10/22/ai-hiring-face-scanning-algorithm-increasingly-decides-whether-you-deserve-job/, last accessed on 12. 10. 2021.

HLEG – High-Level Expert Group on Artificial Intelligence (2019): Ethics guidelines for trustworthy AI. Available online at https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60419, last accessed on 28. 09. 2021.

Köchling, Alina; Wehner, Marius (2020): Discriminated by an algorithm. A systematic review of discrimination and fairness by algorithmic decision-making in the context of HR recruitment and HR development. In: Business Research 13 (3), pp. 795–848. https://doi.org/10.1007/s40685-020-00134-w

Lyell, David; Coiera, Enrico (2016): Automation bias and verification complexity. A systematic review. In: Journal of the American Medical Informatics Association 24 (2), pp. 424–431. https://doi.org/10.1093/jamia/ocw105

O'Neil, Cathy (2016): Weapons of math destruction. How big data increases inequality and threatens democracy. New York, NY: Crown.

Obermeyer, Ziad; Powers, Brian; Vogeli, Christine; Mullainathan, Sendhil (2019): Dissecting racial bias in an algorithm used to manage the health of populations. In: Science (336), pp. 447–453. https://doi.org/10.1126/science.aax2342

Parasuraman, Raja; Manzey, Dietrich (2010): Complacency and bias in human use of automation. An attentional integration. In: The Journal of the Human Factors and Ergonomics Society 52 (3), pp. 381–410. https://doi.org/10.1177/0018720810376055

Selbst, Andrew; Boyd, Danah; Friedler, Sorelle; Venkatasubramanian, Suresh; Vertesi, Janet (2019): Fairness and abstraction in sociotechnical systems. In: Association for Computing Machinery New York, NY (ed.): FAT* '19, Proceedings of the conference on fairness, accountability, and transparency, pp. 59–68. https://doi.org/10.1145/3287560.3287598

Simon, Judith; Wong, Pak-Hang; Rieder, Gernot (2020): Algorithmic bias and the value sensitive design approach. In: Internet Policy Review (9) 4, p. 1–16. https://doi.org/10.14763/2020.4.1534

Strauß, Stefan (2018): From big data to deep learning. A leap towards strong AI or 'intelligentia obscura'? In: Big Data and Cognitive Computing 2 (3), pp. 1–19. https://doi.org/10.3390/bdcc2030016

Strauß, Stefan (2021): Deep automation bias. How to tackle a wicked problem of AI? In: Big Data and Cognitive Computing 5 (2), pp. 1–14. https://doi.org/10.3390/bdcc5020018

Tsamados, Andreas et al. (2020): The ethics of algorithms. Key problems and solutions. Available online at SSRN's eLibrary. https://doi.org/10.2139/ssrn.3662302

Tsoukiàs, Alexis (2020): Social responsibility of algorithms. An overview. Available online at https://arxiv.org/pdf/2012.03319.pdf, last accessed on 12. 10. 2021.

Wieringa, Maranke (2020): What to account for when accounting for algorithms. A systematic literature review on algorithmic accountability. In: Association for Computing Machinery New York, NY (ed.): FAT* '19, Proceedings of the conference on fairness, accountability, and transparency, pp. 1–18. https://doi.org/10.1145/3351095.3372833

DR. STEFAN STRAUSS

is senior scientist at the Institute of Technology Assessment (ITA) at the Austrian Academy of Sciences. His main research focus is on the interplay between technology and society, governance of socio-technical systems and the question how digitization affects social practices, human rights, policy and value systems.

OPEN ACCESS oekom

RESEARCH ARTICLE

Künstliche Intelligenz parlamentarisch (mit)gestalten

Vergangene technische Zukünfte in den Berichten der Enquete-Kommissionen des Deutschen Bundestags

Christian Vater, *Akademie der Wissenschaften und der Literatur Mainz, Digitale Akademie, Geschwister-Scholl-Straße 2, 55131 Mainz, DE (christian.vater@adwmainz.de)* *0000-0003-1367-8489*
Eckhard Geitz, *Institut für Soziologie, Albert-Ludwigs-Universität Freiburg, Freiburg, DE (eckhard.geitz@soziologie.uni-freiburg.de)*

Zusammenfassung • Seit 40 Jahren beschäftigen sich Enquete-Kommissionen des Deutschen Bundestages in ihren Berichten mit Fragen der Auswirkungen „neuer Informations- und Kommunikationstechnologien" auf die bundesdeutsche Gesellschaft, worunter bereits frühzeitig auch ‚künstliche Intelligenz' fällt. Im vorliegenden Aufsatz wird ein erster Überblick über die Entwicklung der inhaltlichen Schwerpunkte wie auch der Präsentationsform in diesem Quellentyp gegeben dessen Aufkommen in den 1980er-Jahren mit dem Anfang der parlamentarischen Technikfolgenabschätzung zusammenfällt. Unsere Forschungsergebnisse zeigen unter anderem, dass aktuelle Berichte weniger konkrete Handlungsempfehlungen an Parlamentarier:innen geben als frühere und somit politische (Mit-)Gestaltungsansprüche weniger stark betonen.

Parliamentarian (co-)design of artificial intelligence.
Past technical futures in the reports of the German Bundestag's committees of inquiry

Abstract • *For 40 years, committees of inquiry of the German Bundestag have been dealing in their reports with questions of the effects of "new information and communication technologies" on German society, including artificial intelligence at an early stage. This article gives an initial overview of the development of the main focus areas and the form of presentation in this source type, whose emergence in the 1980s coincides with the beginning of parliamentary technology assessment. Our research findings show that current reports give less concrete policy recommendations to parliamentarians than earlier ones, thus placing less emphasis on political (co-)design.*

https://doi.org/10.14512/tatup.30.3.50
Received: Jun. 27, 2021; revised version accepted: Oct. 19, 2021; published online: Dec. 20, 2021 (peer review)

Keywords • *artificial intelligence, enquete, German Federal Parliament, technology assessment, information technology*

Einleitung

In Zeiten der Veränderung, in denen die Anzeichen eines Übergangs so zahlreich sind, dass allgemein anerkannt von einer Epochenschwelle gesprochen werden kann, steigt der individuelle wie der kollektive Orientierungsbedarf. Diesem kann auf verschiedene, prinzipiell gleichwertige Weisen gedient werden: Man kann sich in der Gegenwart einer Sache versichern und mit Hilfe von Umfragen und Untersuchungen synchron eine Lage kartieren und so die eigene Position feststellen und perspektivieren. Man kann über die Zukunft spekulieren und in die interessante, aber unzugängliche Zeitspanne zwischen dem morgigen Tag und möglichen ‚langen Wellen' mathematisch modellierte Prognosen oder heuristisch gegriffene Szenarien einschreiben, mit dem typischen unsicheren Geltungsanspruch (Koselleck 2003). Man kann aber auch das Archiv konsultieren und das Labyrinth der Akten und Aufzeichnungen betreten, um die Geltungsansprüche der als selbstverständlich gesetzten Vorannahmen der gegenwärtigen Debatte zu befragen: Sind etwa Phänomene wie Digitalisierung und künstliche Intelligenz wirklich so neu, wie es in aktuellen Diskursen bisweilen nahegelegt wird? Sind die damit verknüpften Herausforderungen tatsächlich so dringend, wie sie dargestellt werden? Sind die Vorschläge gegenwärtiger Berater:innen und Expert:innen wirklich singulär oder stehen sie in Traditionen? Oder: Gibt es Traditionslinien, die vergessen oder gebrochen wurden? Lassen sich mediale Spuren im kulturellen Gedächtnis aufzeigen (Assmann und Assmann 1994)?

Der vorliegende Aufsatz möchte aufzeigen, dass zumindest eine Textgruppe aus dem Archiv des Deutschen Bundestags –

dem kollektiven Gedächtnisspeicher der entscheidungstreibenden ‚Herzkammer' unseres Parlaments – zur Annäherung an eine Antwort auf die soeben gestellten Fragen unsere Aufmerksamkeit verdient: die Berichte der fachlich einschlägigen Enquete-Kommissionen zu neuen Informations- und Kommunikationstechniken. An jedem dieser Zeitdokumente könnte man innehalten und das darin verwobene, sich aus Wissenschaft, Wirtschaft und Gesellschaft speisende Wissen sortieren und freilegen. Jeder Bericht könnte wiederum zentraler Knotenpunkt eines Akteur-Netzwerks beteiligter Menschen und Nicht-Menschen werden. Diese Detailzeichnung kann hier allerdings aus pragmatischen Gründen nicht vorgenommen werden. Vielmehr geht es darum, in einem ersten Schritt und nach kursorischer Zusammenschau zu verdeutlichen, dass sich weitere Folgeschritte ins Archiv lohnen könnten – um gleichermaßen methodisch und inhaltlich von den handelnden Technologieerklärer:innen der Vergangenheit zu lernen.

Quellen und methodische Überlegungen

Die verarbeiteten Berichte zeigen als Zeitdokumente, neben allen historischen Feinheiten der Entwicklung von Technik(en) und ihrer Debatte(n) auch die Entwicklung der Technikfolgenabschätzung in Deutschland auf: von sachbezogenen Klärungsaufträgen an kleine Gruppen moderierender Parlamentarier:innen über die Institutionalisierung der Technikfolgenabschätzung bis hin zur aufwendig organisierten Großdebatte. Im vorliegenden Text konzentrieren wir uns vornehmlich auf die ersten beiden Berichte von 1983 und 1990, da diese am Anfang der Entwicklung der Technikfolgenabschätzung in Deutschland stehen. Lediglich kurze Erwähnung finden die Berichte von 1998 (Deutschland auf dem Weg in die Informationsgesellschaft), 2013 (Digitale Gesellschaft) und 2020, der sich explizit dem Topos ‚künstliche Intelligenz' widmet.

Ein methodischer Befund unserer Forschung soll nicht unerwähnt bleiben: Der Fokus auf die Quellengruppe der Enquete-Kommissions-Berichte – eine Textart mit eigenen Genese- und Rezeptionsbedingungen – hat den Blick auf diachrone Entwicklungen eines ‚weiten' Sachfeldes geschärft. Statt also ‚semasiologisch' entlang von Schlagwörtern zu suchen, zu finden und zu ordnen, haben wir ‚onomasiologisch' den Zusammenhang greifen können, der „Informations- und Kommunikationstechnik", „Expertensysteme", „Neue Medien", „Digitale Gesellschaft" und „künstliche Intelligenz" jenseits aller Begriffsarbeit verbindet.

Auf dem Weg ins tiefe Archiv der Technologie-Debatten

Anfang der 1980er-Jahre beginnt die Öffentlichkeit, sich für die Veränderungen ihrer Kommunikationsumgebung und ihrer medialen Kanäle in höchstem Maße zu interessieren. Nicht nur treten mit Marshal McLuhans Kurz-Auftritt in Woody Alans ‚Annie Hall' 1977 Medienwissenschaftler:innen als Persönlichkeiten endgültig in den Populärdiskurs ein, auch das Nachdenken über die Wechselwirkung zwischen medialem Wandel und gesellschaftlicher Wirklichkeit wurde entlang der Rolle des Fernsehens und vor der Anschauung des ‚Globalen Dorfes' diskutiert. 1979 hatte der französische Philosoph Jean-François Lyotard für den Universitätsrat Québecs einen Bericht über den Wandel des Wissens in einer technologisch veränderten Welt verfasst. Er stellt fest, dass – auch wenn es „unklug [sei], der Futurologie exzessiven Kredit einzuräumen" – man nicht umhinkäme, festzuhalten, dass die „Auswirkungen […] technologischer Transformation auf das Wissen […] erheblich" (Lyotard 2019, S. 23 f.) seien. Die technologische Transformation werde getrieben von Kybernetik, Kommunikation und Telematik, von ‚modernen Algorithmen', Informatik, Computern, Datenbanken und der „Perfektionierung ‚intelligenter' Terminals" (Lyotard 2019, S. 30). Dieses ‚postmoderne Wissen' löste weltweite Diskussionen aus, die auch in Deutschland anbrandeten: Am 9. April 1981 erhält die Enquete-Kommission „Neue Informations- und Kommunikationstechniken" vom Deutschen Bundestag in Bonn ihren Arbeitsauftrag. Mit diesem Arbeitsauftrag beginnt eine (mittellange) Geschichte der parlamentarischen Beratung von und Debatte über neue Technologien und Techniken, die wir heute unter den Schlagworten „Digitalisierung" und „künstliche Intelligenz" weiterführen. Entlang der einzelnen Berichte kann man sie auf ihrem Zeitstrahl folgendermaßen gliedern:

1983 sehen einige Experten deutlich, dass tiefgreifende Entwicklungen eintreten werden und es wird der Bedarf an Aufklärung der technischen Grundlagen im Dienst einer informierten Forschungs-, Wirtschafts-, Bildungs- und Sozialpolitik angemeldet. Fragen nach Netzinfrastruktur, Hardwareentwicklung und Erhebungsmethoden einer wissenschaftlichen Technikfolgenabschätzung stehen im Mittelpunkt (Deutscher Bundestag 1983). Die Berichtslegung fällt in ein für Deutschland folgenschweres Jahr. Auf dem Höhepunkt der Friedensbewegung beginnt die Kanzlerschaft Helmut Kohls, dessen Koalitionsregierung weitreichende marktliberale Reformen auf den Weg bringt.

1990, im Jahr der Wiedervereinigung und ein Jahr nachdem mit der Privatisierung der Deutschen Bundespost begonnen wurde, konkretisieren sich die Anwendungsfälle und Orientierungsansprüche, technologische Durchbrüche werden von Fachleuten genauso allgemeinverständlich aufbereitet wie deren allgemeine Grundlagen und vermuteten Folgen, wobei sich die Darstellung in klaren Handlungsoptionsangeboten verdichtet. Der Bericht ist zudem infographisch auf höchstem Niveau durchgestaltet. Ein Schwerpunkt der Fragen liegt auf Softwareentwicklung und -einsatz (Deutscher Bundestag 1990).

1998, zum Ende der Ära Kohl, wird die gesellschaftliche Aneignung der technischen Innovation verhandelt, strategische Abstraktionen im Dienst konkreter Politik- und Unternehmensberatung greifen Raum. (Technik-)soziologische Fragestellungen werden formuliert. Der Bundestag pflegt eine Webseite und stellt Infomaterial zur Verfügung (Deutscher Bundestag 1998).

2013 markieren konkrete rechtliche Klärungsbedarfe die technikbedingten Reibungen und Nutzungsansprüche der Zivil-

gesellschaft mit den Unternehmungen einer global verschalteten, digitalen und multimedialen Kommunikation. Der Bericht selbst ist hinsichtlich Schreibtechnik, multimedialer Netzpräsentation und Bürgerbeteiligungsangeboten ein Medienexperiment, inklusive webarchivierter Videostreams einiger Anhörungen und Sitzungen sowie abgedruckter qualitativer Expert:inneninterviews (Deutscher Bundestag 2013).

2020 wird eine Geschichte möglicher Zukünfte verfasst – nun in groben Linien entkoppelt von konkreten technischen Materialitäten und Fragen, ihrer (Er-)Klärung oder Vermittlung – und im Modus der gemeinsamen Entwicklung einer ziel- und nicht sachstandsorientierten Narration werden technikbasierte Hoffnungen und Wünsche eines breiten Spektrums von Interessensgruppen gesammelt. ‚Menschenzentrierte KI' und ‚KI Made in Germany' sind die Brennpunkte und Zielvorstellungen eines Aushandlungsprozesses, an dessen Ende das mit Abstand umfangreichste Dokument dieser Reihe als Bundesdrucksache verfertigt wird. Das infographische Bildvokabular von Unternehmensberater:innen ist mit den Kreuztabellen der SWOT-Analyse (englisch für Stärken, Schwächen, Chancen, Risiken) genauso parlamentsfähig geworden wie das überreiche Verzeichnis ephemerer Onlinequellen, welches das wissenschaftliche Literaturverzeichnis abgelöst hat. Ein Onlinebefragungsportal zur niedrigschwelligen Bürgerbeteiligung begleitet die Kommissionsarbeit, wird allerdings nur in geringem Umfang genutzt (Deutscher Bundestag 2020).

Im Gegensatz zum letzten vorliegenden Enquete-Bericht beginnt die Reihe mit dem aus dem Jahr 1983 sehr konkret: mit Satelliten und Glasfaserkabeln, mit den materiellen Schaltungen und den erwartbaren Anwendungen der Informations- und Telekommunikationstechnologie.

Im April 1981 erhielt die Enquete-Kommission ‚Neue Informations- und Kommunikationstechniken' vom Deutschen Bundestag in Bonn ihren Arbeitsauftrag.

1983 – Hardware. Technische Präzision und Bemühen um technologische Aufklärung: Vernetzte Informations- und Kommunikationstechnik als technologische Bedingung. Techniker:innen vermitteln

Die parlamentarische Auf- und Abklärung der Möglichkeiten und Grenzen digitaler Neuer Informations- und Kommunikationstechnik (IuK) muss ihren ersten, vielversprechenden Schritt leider verstolpern: Mitten während des Arbeitsprozesses werden nach dem Bruch der Regierungskoalition außerplanmäßige Wahlen ausgerufen, ein Abschlussbericht der am 9. April 1981 eingerichteten Enquete-Kommission kann nicht mehr ausgearbeitet werden. Gleichwohl ergeben sich aus dem dezidierten Transfer technologischer (Un-)Möglichkeiten auf die relevanten politischen, sozialen und kulturellen Bereiche der BRD zu Beginn der 1980er-Jahre umfangreiche Desiderata hinsichtlich steuernder und regulierender technopolitischer Maßnahmen – etwa den Ausbau des Glasfasernetzes bis Ende der 1980er-Jahre, siehe den abschließenden Diskussionsstand der Enquete-Kommission (Deutscher Bundestag 1983, S. 9–13 b). Ob eben dieser Steuerungsanspruch der Grund dafür war, dass kein weiteres Interesse an seiner Fertigstellung bestand, steht dahin. Außerordentlich vielversprechend und gehaltvoll ist jedenfalls der Zwischenbericht, der am 28. 03. 1983 veröffentlicht wird. Der Zwischenbericht beginnt mit zwei Thesen, auf die sich die Mitglieder ausdrücklich einigen konnten:

> „These 1: Die Entwicklung neuer Technologien, einschließlich der IuK-Techniken und deren Nutzung, ist von den jeweils gegebenen politischen, ökonomischen, sozialen und rechtlichen Bedingungen abhängig." (Deutscher Bundestag 1983, S. 8)

> „These 2: Die Analyse der technologischen, ökonomischen, sozialen und rechtlichen Aspekte der Kommunikationstechnologien und die Folgenabschätzung treffen auf Schwierigkeiten bei der Erfassung der Tatsachenbasis, auf Unsicherheiten der theoretischen Annahmen über Wirkungszusammenhänge sowie auf Kontroversen bei der Bestimmung von Kriterien für die Auswahl entscheidungserheblicher Gegenstände und die Bewertung der Befunde." (Deutscher Bundestag 1983, S. 8)

Wenn also, so die Kommission, neue Technologien entwickelt und in einem zweiten Schritt auch genutzt werden sollen, hat dies komplexe gesellschaftliche und kulturelle Voraussetzungen. Außerdem seien zur Zeit der Abfassung dieses Zwischenberichtes sowohl praktisches Technikverständnis und theoretisches Systemmodell als auch Auswahl- und Bewertungskriterien von Sachverhalten und Befunden unzureichend.

Ausgehend von einer detaillierten Darstellung des technischen Aufbaus der Hardware der Kommunikationsnetze (Kapitel 1) geht der Zwischenbericht der Kommission über den aktuellen Stand von Hersteller- und Anwenderbranchen sowie einem kurzen Ausblick auf Verbraucher (Kapitel 2) über zur mit 100 Seiten umfassenden Abschätzung der Auswirkungen auf Wirtschaft und Gesellschaft (Kapitel 3). Es folgt abschließend die Behandlung besonders hervorgehobener Teilaspekte (Kapitel 4), wozu Normung, Forschungsförderung, Datenschutz, Verbraucherschutz und Urheberrecht gehören. Der detaillierte Text

selbst bettet Fachausdrücke der Verständlichkeit halber in zugängliche und konkrete Anwendungszusammenhänge ein (z. B. hinsichtlich der Schlüsseltechnologie ISDN, Deutscher Bundestag 1983, S. 30) und liefert graphisch abgesetzte und kognitiv leicht zugänglich gestaltete Kurzzusammenfassungen (z. B. hinsichtlich relevanter technischer Innovationen, Deutscher Bundestag 1983, S. 30).

Einige der diskutierten Problemstellungen wirken heute allzu bekannt, so etwa die Frage, ob „Kreativität, Intelligenz, Empathie, Bindungsfähigkeit, Solidarität und Verantwortungsbereitschaft" durch ‚neue Medien' leiden würden (Deutscher Bundestag 1983, S. 154). Andere Kapitel, etwa „3. Bedeutung und Auswirkungen der IuK-Techniken in Wirtschaft und Gesellschaft" wirken wiederum brandaktuell und haben sogar noch immer eine utopische Komponente, wenn es um die „Demokratisierung des Informationszugriffs" geht, und zwar „für alle Bürger" (Deutscher Bundestag 1983, S. 166).

Zur Gestaltung des Produktiveinsatzes der neuen Technologien findet sich im Infokasten 2 des Unterkapitels 3.2.3 „Arbeitsplatz, Heimarbeit, Gesundheitsschutz, Mitbestimmung" (Deutscher Bundestag 1983, S. 106 f.) eine heute ungewohnte Haltung partizipativer betrieblicher Mitbestimmung:

> „Die Nutzung der IuK-Technologien (Geräte, Systeme, Dienste) zu arbeitsplatzrelevanten Innovationen ermöglicht eine starke Erhöhung (Ausweitung) des sozialorganisatorischen Gestaltungsspielraums hinsichtlich der Arbeitsbedingungen und der Qualifikationsstrukturen. Welche der möglichen Gestaltungsoptionen sich durchsetzen können, hängt weniger von den neu eingesetzten Techniken und den daraus folgenden technischen Restriktionen ab, sondern vielmehr von der organisatorischen Konzeption des betrieblichen Ablaufes, der Lern- und Anpassungsfähigkeit aller am Arbeitsprozeß Beteiligten, sowie der politischen und vertraglichen Durchsetzungskraft der Tarifvertragsparteien."

Die hier hinterlegte Einordnung des Verhältnisses neuer Lehr-Technologien (heute E-Learning) zur Mitbestimmung und Berufsfachlichkeit ist selbst in den Pandemie- und Digitalisierungsjahren 2020 und 2021 noch kein Allgemeingut.

Die Sphäre des Kulturellen wird gleichzeitig als genauso betroffen gedacht, da ihr eine grundsätzliche Medialität zugeschrieben wird, so im Infokasten 1 im Unterkapitel 3.2.4 „Ausbildung, Bildung, Wissenschaft, Kultur" (Deutscher Bundestag 1983, S. 122):

> „Wesentliche Grundlage der Kultur in einem weiteren – über den engen Bereich der Künste und der Kunstpflege hinausweisenden Sinn – ist die Kommunikation. Die neuen IuK-Techniken erweitern die Möglichkeiten technisch vermittelter Kommunikation. Sie können damit nicht nur Veränderungen im Denken, Erleben und Verhalten des einzelnen, sondern auch im allgemeinen gesellschaftlichen Kulturprozeß bewirken."

Der Darstellung folgen ein Anmerkungsteil mit namentlichen Kommentaren und Änderungsvorschlägen auf Stand der Diskussion mit frühzeitiger Angabe (Deutscher Bundestag 1983, S. 217–237) sowie ein brauchbares Glossar. Es wurden sieben Sachverständige berufen, drei Anhörungen veranstaltet, vier Unterkommissionen eingerichtet, 150 schriftliche Stellungsnamen eingearbeitet, Informationsreisen ins Ausland getätigt und als Partner das Deutsche Institut für Wirtschaftsforschung, das Heinrich-Hertz-Institut, und die PROGNOS AG eingebunden.

Gerade die frühzeitige Beschäftigung mit einer Technik soll das Parlament in die Lage versetzen, sie zum Wohl des Ganzen zu gestalten, statt nachträglich unerwünschte Folgen zu beseitigen.

1990 – Software. Zentrale Innovationsfelder werden abgesteckt und kartiert: ‚Expertensysteme' in industrieller Fertigung und Medizintechnik im Detail. Entwickler:innen wollen gestalten

Schon der Bundestag der zehnten Wahlperiode hatte eine Enquete-Kommission „Technikfolgen-Abschätzung und -Bewertung" eingerichtet. Dieser Ansatz war neu, da erst 1989 der erste Bericht dieser Kommission vorgelegt wurde, der Technikfolgenabschätzung (TA) überhaupt als Arbeitsfeld ordnete – das Anwendungsfeld der ‚Expertensysteme' musste als nur teilbehandelt an die Nachfolge in der elften Wahlperiode übergeben werden (Deutscher Bundestag 1990, S. 9). Hier wurden nun zwei methodische Zugänge der TA exemplarisch erprobt: Ein „problem**in**duzierter" mittels des Sachfeldes „Alternativen landwirtschaftlicher Produktionsweisen" und ein „technikzentrierter" mittels der „Expertensysteme". Das Parlament sollte anhand des Vergleichs der Ergebnisse beurteilen, welcher der erprobten methodischen Zugänge angemessener oder brauchbarer für die Bedarfe parlamentarischer Praxis sei. „Außerdem", so wurde argumentiert, könne „man mit dieser Untersuchung einen Beitrag zur ‚Entmystifizierung' der künstlichen Intelligenz leisten." (Deutscher Bundestag 1990, S. 9) Der Bericht beginnt so plastisch wie deutlich:

> „‚Ist der Mensch ein Roboter?', ‚Sind Computer lebendig?', dies sind die Titel von nur zwei der vielen populärwissen-

> schaftlichen Bücher, die in den letzten Jahren zur ‚Künstlichen Intelligenz' erschienen sind. Dabei ist die Bezeichnung ‚Künstliche Intelligenz' bereits mehr als dreißig Jahre alt, und die Ideen, die den damit bezeichneten Forschungsarbeiten zugrunde liegen, beschäftigen die Philosophen schon seit der Antike." (Deutscher Bundestag 1990, S. 7)

Nach dieser dezidiert philosophischen Einordnung in einen historischen Prozess, der sich mindestens bis ins klassische Athen zurückverfolgen lasse, folgt ein abklärender, versachlichender und gleichzeitig phantastisch fabulierender, einhegender Befund:

> „Um es gleich vorwegzunehmen: dieser Bericht kann auf die Frage, ob der Mensch eine Maschine sei oder ob uns eines Tages Maschinen als Träger der höchsten Intelligenz beerben werden, keine Antwort geben. Untersucht werden ‚Expertensysteme', Computerprogramme, die aus der Forschung zur ‚Künstlichen Intelligenz' hervorgegangen sind." (Deutscher Bundestag 1990, S. 7)

Die Kommission will also kein Urteil vorwegnehmen, sondern möglichst zugänglich und verständlich Material für das Parlament vorbereiten, und zwar rechtzeitig, um politische Gestaltung zu ermöglichen. Diese soll nach engagierter und kritischer Diskussion „über den Bericht und insbesondere über die von ihr zusammengetragenen Handlungsoptionen" (Deutscher Bundestag 1990, S. 3) durch die Parlamentarier:innen stattfinden. Das Ziel, Handlungs- und Gestaltungsoptionen aufzuzeigen, verdeutlicht sich in ausführlichen Beschreibungen und Ausführungen im Fließtext, durchbrochen von 19 Infokästen und 19 Einfassungen mit klar ausgearbeiteten Optionen für Handlungsfelder, Übersichtstabellen erleichtern zudem den Zugang. Die Behandlung des Themas beginnt bei Geschichte, Begriffsklärung (z. B. ‚künstliche Intelligenz: Eine mißratene Bezeichnung?'), und technischen Grundlagen, behandelt mit dem Einsatz in der Produktion und dem Einsatz in der Medizin zwei epistemisch unterschiedliche Praxisfelder und endet bei Querschnittsthemen wie Risiken (z. B. „Fehlerquellen bei Expertensystemen") oder Datenschutz.

Die Optionen der Handlungsfelder sind jeweils als ganz konkrete parlamentarische Handlungsmöglichkeiten gestaltet. Ein spannendes Beispiel findet sich im Handlungsfeld 16 „Haftung bei Expertensystemen": „Die Bundesregierung auffordern, die ärztliche Haftpflicht auf ihre Anwendbarkeit im Zusammenhang mit der Verwendung von Computerprogrammen (bei fehlerhaftem Einsatz) sowie bei Nichtanwendung zu überprüfen und dabei die Relevanz des Begriffs ‚ärztlicher Kunstfehler' zu erörtern." (Deutscher Bundestag 1990, S. 75) Ein weiteres im Handlungsfeld 18 „Mitbestimmung": „Die Bundesregierung auffordern, eine Untersuchung über Ansätze, Formen und Erfahrungen mit der ‚partizipativen Entwicklung und Gestaltung' informationstechnischer Systeme durchführen zu lassen." (Deutscher Bundestag 1990, S. 84)

Die Kommission von 1990 griff auf Materialien ihrer Vorgängerin zurück, die unter dem Titel „Chancen und Risiken des Einsatzes von Expertensystemen in Produktion, Verwaltung, Handwerk und Medizin" vom Battelle-Institut und dem Institut für Medizinische Informatik und Systemforschung der Gesellschaft für Strahlen- und Umweltforschung geliefert worden waren. Sie beauftragte selbst das Fraunhofer Institut für Arbeitswirtschaft und Organisation (IAO), das Institut für Sozialwissenschaftliche Forschung (ISF) und das Institut für Medizinische Informatik und Systemforschung (medis) der Gesellschaft für Strahlen- und Umweltforschung. Achtzehn Sachverständige wurden gehört, neunzehn weitere um schriftliche Stellungsnamen gebeten. Es folgen neben den erwähnten Tabellen zu „Kästen" und zu „Optionen" (also „Wissen" und „Handlungsaufforderungen") ein auch heute noch belastbares Literaturverzeichnis, ein brauchbares Glossar und ein Verzeichnis der mit dem Bericht verbundenen Kommissionsvorlagen. Die vorbildliche graphische Gestaltung und Aufbereitung dieses Berichtes sollten folgende Berichte nicht wieder erreichen.

Können ‚Technikzukünfte' ohne ‚Technikvergangenheiten' gelingen? – ein (Zwischen-)Fazit

Der aktuelle Enquete-Bericht zur ‚Künstlichen Intelligenz' hat vom Präsidenten des Bundestages einen klaren Auftrag erhalten: Da „die Dynamik der Digitalisierung […] mit der Forschung zur Künstlichen Intelligenz eine neue Dimension erreicht" habe, müsse geklärt werden, „was KI eigentlich bedeute, was sie leisten könne und welche Chancen und Herausforderungen für Staat, Gesellschaft und Recht entstünden." (Deutscher Bundestag 2020, S. 43) Man könnte auch festhalten: Noch immer! Da es sich – im eskalativen Extremfall – nicht nur um Wirtschaft, sondern auch um Waffen handele, wären die Fragen dringlich, „wie diese Entwicklung so gestaltet werden könne, dass KI den Menschen diene". (ibid.) Dass ein nachhaltiger und wohlstandsorientierter Einsatz KI Made in Germany gefördert werden soll, ist jedenfalls Zielvorstellung – auch, wenn Bundestagspräsident Schäuble selbst und bezeichnender Weise von einer „neue[n] Zauberformel des technischen Fortschritts" spricht.

Lässt sich aber ein so großes wie gewichtiges Projekt ohne Blick in die Geschichte – und die Genese der eigenen Grundlagen – belastbar abschließen?

Blickt man nach mehr als 30 Jahren in die hier vorgestellten Berichte, muss man der Technikfolgeabschätzung zubilligen, dass sie hier ausgesprochen treffsicher Technikzukünfte antizipiert hat – im ersten Fall mit Implikationen für politische Steuerung, im zweiten Fall mit klarem Blick für künftige Diskurse über Potenziale und Probleme der künstlichen Intelligenz. Wenn heute die etwas lapidare Rede davon ist, Deutschland habe die Digitalisierung verschlafen, bedarf diese Einordnung einer Richtigstellung. Während etwa das Glasfasernetz mit über 35-jähriger Verspätung allmählich auch im ländlichen Raum ankommt,

war bereits zu Beginn der 1980er-Jahre absehbar, wozu dessen Vernachlässigung führt und was sie verhindert. Technikfolgenabschätzer:innen haben mit den beiden Berichten zwei starke Argumente, mit denen sie zeigen können, dass Technikzukünfte kein Produkt phantasiereicher Spekulationen sein müssen, sondern Szenarien sein können, die politische Steuerung ermöglichen und Versäumnisse dingfest machen. Denn auch die Grundsatzfrage nach der ‚Planung' auf Grundlage verlässlicher und belastbarer Prognosen begleitet die deutschsprachige Technikfolgenabschätzung seit ihren Anfängen, mal als ‚Reizvokabel', mal in Euphorie (Brinkmann 2006, S. 177 ff.).

Eine Rückkehr zum politischen Gestaltungsanspruch der ersten, hier aus dem Archiv gehobenen Berichte halten wir – auch als historiographische Erweiterung eines ‚problemorientierten' Ansatzes – für ratsam. Methodisch sehen wir im hier greifbaren Bemühen um Grundlagenvermittlung bei gleichzeitiger Allgemeinverständlichkeit auf dem technischen Stand der (jeweiligen) Gegenwart ein Vorbild, auch weil alle Mittel des Infographikdesigns in ihrer Übersetzungsleistung genutzt wurden, um der Kommunikation wissenschaftlicher Erkenntnisse in den als partizipativ gedachten politischen Raum zu dienen. Es könnte in einem nächsten Schritt um Mitgestaltung und Mitbestimmung gehen, um die zwei Leitgedanken teilhabender gemeinschaftsorientierter Technologiebesprechung – und um Umnutzung und Weiternutzung, die zwei Leitgedanken emanzipatorischer individueller Aneignungsansprüche. Mit Sicherheit sind jedenfalls die Arbeiten von Enquete-Kommissionen und TA zu künstlicher Intelligenz – als Begriff, als Technik, als Zaubertrick, als Traum – mit dem aktuellen Bericht nicht zu einem befriedigenden Ende gekommen.

Es wird sich also lohnen, die ‚technischen Zukünfte der Vergangenheit', wie sie aus den Berichten der Enquete-Kommissionen des Bundestags hervorgehen, zur Kenntnis zu nehmen, um so der „Struktur solcher ‚Vorstellungen' selbst und ihre[r] realitätsprägenden Funktionen" (Popplow 2020, S. 43) auf die Spur zu kommen – und zwar von den Begriffen über Diskursspuren bis zur Gestaltung ihrer Präsentation und den Blick auf zugrunde liegende Sachverhalte.

Angabe von Finanzierungsquellen

Dieser Forschungsartikel hat keine Förderung erhalten.

Literatur

Assmann, Aleida; Assmann, Jan (1994): Das Gestern im Heute. Medien und soziales Gedächtnis. In: Klaus Merten, Siegfried Schmidt und Siegfried Weischenberg (Hg.): Die Wirklichkeit der Medien. Eine Einführung in die Kommunikationswissenschaft. Wiesbaden: Springer, S. 114–140. https://doi.org/10.1007/978-3-663-09784-6_7

Brinckmann, Andrea (2006): Wissenschaftliche Politikberatung in den 60er Jahren. Die Studiengruppe für Systemforschung, 1958 bis 1975. Berlin: edition sigma. https://doi.org/10.5771/9783845271095

Deutscher Bundestag (1983): Zwischenbericht der Enquete-Kommission „Neue Informations- und Kommunikationstechniken" gemäß Beschluß des Deutschen Bundestages vom 9. April 1981. Bonn: Bundesdrucksache 09/2442. Online verfügbar unter https://dserver.bundestag.de/btd/09/024/0902442.pdf, zuletzt geprüft am 18. 10. 2021.

Deutscher Bundestag (1990): Bericht der Enquete-Kommission „Gestaltung der technischen Entwicklung; Technikfolgen-Abschätzung und -Bewertung" gemäß Beschluß des Deutschen Bundestages vom 5. November 1987. Bonn: Bundesdrucksache 11/7990. Online verfügbar unter http://dipbt.bundestag.de/doc/btd/11/079/1107990.pdf, zuletzt geprüft am 18. 10. 2021.

Deutscher Bundestag (1998): Schlußbericht der Enquete-Kommission Zukunft der Medien in Wirtschaft und Gesellschaft – Deutschlands Weg in die Informationsgesellschaft (eingesetzt durch Beschluß des Deutschen Bundestages vom 7. Dezember 1995) zum Thema Deutschlands Weg in die Informationsgesellschaft. Bonn: Bundesdrucksache 13/11 004. Online verfügbar unter http://dipbt.bundestag.de/doc/btd/13/110/1311004.pdf, zuletzt geprüft am 18. 10. 2021.

Deutscher Bundestag (2013): Schlußbericht der Enquete-Kommission „Internet und digitale Gesellschaft" eingesetzt durch Beschluss des Deutschen Bundestages vom 4. März 2010. Berlin: Bundesdrucksache 17/12 550. Online verfügbar unter https://dserver.bundestag.de/btd/17/125/1712550.pdf, zuletzt geprüft am 18. 10. 2021.

Deutscher Bundestag (2020): Unterrichtung der Enquete-Kommission Künstliche Intelligenz – Gesellschaftliche Verantwortung und wirtschaftliche, soziale und ökologische Potenziale (eingesetzt durch Beschluss des Deutschen Bundestages vom 26. Juni 2018). Bericht der Enquete-Kommission Künstliche Intelligenz – Gesellschaftliche Verantwortung und wirtschaftliche, soziale und ökologische Potenziale. Berlin: Bundesdrucksache 19/23 700. Online verfügbar unter https://dserver.bundestag.de/btd/19/237/1923700.pdf, zuletzt geprüft am 18. 10. 2021.

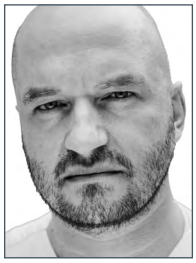
Koselleck, Reinhart (2003): Zeitschichten. Studien zur Historik. Frankfurt am Main: Suhrkamp.

Lyotard, Jean-François (2019): Das postmoderne Wissen. Ein Bericht. Berlin: Passagen.

Popplow, Marcus (2020): Zur Erforschung von Technikzukünften aus technikhistorischer Perspektive. In: Paulina Dobroç und Andie Rothenhäusler (Hg.): 2000 revisited. Rückblick auf die Zukunft. Karlsruhe: KITopen, S. 41–58. https://doi.org/10.5445/KSP/1000117728

CHRISTIAN VATER

ist Wissenschaftlicher Mitarbeiter an der Digitalen Akademie der Akademie der Wissenschaften und der Literatur Mainz sowie Lehrbeauftragter am Department für Geschichte des Instituts für Technikzukünfte (ITZ) des Karlsruher Instituts für Technologie (KIT). Als Wissenschaftstheoretiker und Technikphilosoph beforscht er die Wissensgeschichte der künstlichen Intelligenz.



ECKHARD GEITZ

ist Leiter des Bildungsinstituts im Gesundheitswesen (BIG) in Essen und arbeitet am Institut für Soziologie der Universität Freiburg an einem Promotionsprojekt zur Digitalisierung der Psychiatrie.

OPEN ACCESS oekom

RESEARCH ARTICLE

Rechtliche Rahmenbedingungen für KI-Systeme

Immanente Herausforderungen und mögliche Lösungen durch Control by Design

Thomas Wilmer, *Institut für Informationsrecht, Hochschule Darmstadt, Haardtring 100, 64295 Darmstadt, DE (thomas.wilmer@h-da.de)*

Zusammenfassung • Der Autor stellt die komplexen rechtlichen Rahmenbedingungen für künstliche Intelligenz aus den Bereichen des Dateneigentums, des Datenschutzes und des Urheberrechts dar und erläutert, weswegen die unübersichtliche Rechtslage nicht dazu geeignet ist, auf Rechtssicherheit durch neue Regulierungen auf EU- oder nationaler Ebene zu hoffen. Stattdessen werden ein Regulierungsrahmen für zulässige Vertragsvereinbarungen sowie dazu passende Control by Design-Technikeinstellungen vorgeschlagen.

Legal framework conditions for AI systems. *Immanent challenges and possible solutions (control by design)*

Abstract • *The author presents the complex legal framework for artificial intelligence in data ownership, data protection, and copyright and explains why the confusing legal situation is not suitable for hoping for legal certainty through new regulations at the EU or national level. Instead, he proposes a regulatory framework for permissible contractual agreements and matching control-by-design settings.*

Keywords • *data ownership, data governance act, copyright, artificial intelligence*

Künstliche Intelligenz (KI) stellt aus rechtlicher Sicht verschiedene Herausforderungen an die Einordnung in das vorhandene Regulierungssystem: KI kann als Software mit vorgefertigten Programm-Elementen und von der Software selbst geschaffenen neuen Programmanteilen verstanden werden, was die Frage nach dem Geistigen Eigentum an den immateriellen Werten stellt, welche von der KI geschaffen werden. Daneben basiert KI immer auch auf einem (dann fortlaufend aktualisierten) Datensatz, welcher in aller Regel der Manifestation menschlicher oder maschineller Erfahrungswerte entspricht. Zu klären ist, wem dieser Datensatz zusteht. Soweit es sich hierbei um personenbezogene Daten handelt, gelten innerhalb der Europäischen Union strikte Regelungen zur Verwendbarkeit dieser Daten. Insgesamt stellt sich Frage nach Folgen des KI-Einsatzes für die Gesellschaft und nach der Transparenz der zugrunde gelegten oder neu geschaffenen Algorithmen sowie der Kontrolle der Einsatzbedingungen.

Bei kritischer Betrachtung der gängigen Geschäftsmodelle kann KI daneben auch als Mittel dienen, um Daten und Know-how der KI-Nutzer zu zentralisieren und dann wiederum allen zur Verfügung zu stellen, wobei die Wertschöpfung bei den KI-Betreibern verbleibt.

KI steht aktuell im Spannungsfeld zahlreicher neuer Regulierungsversuche auf nationaler und europäischer Ebene.

KI steht aktuell im Spannungsfeld zahlreicher neuer Regulierungsversuche auf nationaler und europäischer Ebene. Neben gesetzlichen Begrenzungen der Einsatzbereiche der KI und der Regelung der Verfügungsgewalt über Datensätze sollte verstärkt auch an die zentrale Frage gedacht werden, welche Möglichkeiten es gibt, vertragliche Vereinbarungen mit Endnutzern zur Nutzung der KI und Übertragung von Daten zu regulieren. Solche vertraglichen Vorgaben könnten dann in das Design der KI-Systeme einfließen und den Rahmen für eine Control by Design-Regelung bilden. Diese könnte eine übergeordnete datenschutzfreundliche Gesamteinstellung sein, welche sich an den Regelungen des Art. 25 EU-Datenschutzgrundverordnung (DSGVO) für Privacy-by-Design (datenschutzfreundliche Systemkonzep-

https://doi.org/10.14512/tatup.30.3.56
Received: Jun. 22, 2021; revised version accepted: Oct. 21, 2021; published online: Dec. 20, 2021 (peer review)

tion) und den Neuregelungen des Telekommunikation-Telemedien-Datenschutzgesetzes (TTDSG) orientiert.

Die Bedeutung der KI für Gesellschaft und Wirtschaft findet ihren Niederschlag in internationalen Bemühungen, die Entwicklung zu analysieren und zu regulieren. Nach Auffassung der deutschen Datenethikkommission (Datenethikkommission 2019) ist eine möglichst europäische Regulierungsreichweite analog zu den datenschutzrechtlichen Regelungen anzustreben. Das europäische Weißbuch zur künstlichen Intelligenz – ein europäisches Konzept für Exzellenz und Vertrauen fordert „angesichts der erheblichen Auswirkungen, die KI auf unsere Gesellschaft und die notwendige Vertrauensbildung haben kann […], dass die europäische KI auf unseren Werten und Grundrechten wie Menschenwürde und Schutz der Privatsphäre fußt" (Europäische Kommission 2020a, S. 2).

Umsetzungsbedarf

Fraglich ist, inwiefern diese Ziele in einer funktionsfähigen, den Interessen der Betroffenen und der Wirtschaft angemessenen Form berücksichtigt werden können. Aus juristischer Sicht stellen sich eine ganze Reihe von Fragen aus verschiedenen Rechtsgebieten an den Einsatz der KI, einschließlich sämtlicher unscharf definierter Formen smarter und selbstlernender Produkte oder des Einsatzes sogenannter ‚Legal Tech' (Herberger 2018). Es bleibt zu klären, ob die neuen Regulierungsansätze und Absichtserklärungen mit bestehenden rechtlichen Rahmenbedingungen umzusetzen sind oder ob es einer grundsätzlich neuen Gesetzgebung bedarf, welche die Zulässigkeit des KI-Einsatzes in bestimmten Einsatzfeldern (etwa der Personal- oder Einstellungsbeurteilung oder des Kampfdrohneneinsatzes) oder in bestimmten Auswertungsbereichen (Big Data/gläserner Bürger) beschränkt. Anzustreben wäre eine homogene Regulierung sowohl der Fragen der Inhaberschaft der von der KI geschaffenen Softwareroutinen, der Zuordnung der ausgewerteten und erzeugten Datensätze, der Begrenzung der Einsatzgebiete und der Haftung für die Einsatzfolgen.

Diese Regelung müsste sich zugleich in KI-Systemen transparent abbilden lassen, so wie dies etwa beim Risikomanagement im Datenschutzbereich nach der DSGVO vorgesehen ist. Andernfalls droht eine Zersplitterung der gesetzlichen Lage, welche es den Betroffenen der KI-Nutzung erschwert, überhaupt einen Überblick über die KI-Nutzung zu erhalten. Je mehr Einzelregulierungen vorhanden sind, welche umfangreiche Informationspflichten enthalten, umso größer ist die Gefahr der sogenannten ‚Consent Fatigue' (Rauer und Ettig 2021): Nutzer verlieren das Interesse an der Wahrnehmung der zahllosen Informationen und sind bereit, Nutzungen zu akzeptieren, ohne überhaupt noch zu lesen, welchen Inhalt die Informationen haben, selbst wenn es eine transparente Zusammenfassung geben sollte. Als Musterbeispiel dafür können die ‚Cookie-Pop-Ups' gelten, die von vielen Nutzern nicht mehr gelesen, sondern nur noch akzeptiert werden. Nach einer Statista-Umfrage vom 04. 06. 2020 gaben 41 Prozent der Befragten in Deutschland an, sich grundsätzlich die Inhalte der Cookie-Hinweise nicht durchzulesen und einfach auf ‚Okay' oder ‚Cookies akzeptieren' zu klicken (Statista Research Department 2020).

Im Folgenden soll vor der Darstellung der auf deutscher und EU-Ebene geplanten Regelungen aufgezeigt werden, welche rechtlichen Rahmenbedingungen grundsätzlich zu beachten sind, die sich mit den Rechten an den durch KI betroffenen Daten und der zugrundeliegenden Software befassen.

Fragen des Geistigen Eigentums

Geistiges Eigentum an den in Programm- und Datenform vorliegenden Ergebnissen der KI wird vor allem an urheberrechtlichen Fragen auszurichten sein. Urheberrechtlich ist zu klären, wer der Schöpfer der Ergebnisse der KI ist, sowohl was neu generierten Programmcode als auch Datensätze betrifft. Handelt es sich um den Urheber des Programms, da dieser typischerweise bereits vorgefertigte Routinen in die KI implementiert hat, welche dann – ebenfalls nach vom Urheber vorgegebenen Algorithmen – zusammengesetzt werden (siehe Abb. 1)? Ist das Ergebnis der KI daher nichts weiter als eine verlängerte Schöpfung durch einen Programmierer? Falls dies so wäre, würde sich jedoch die Frage stellen, ob KI dann selbst eine ausreichende Kreativität beinhaltet und so wiederum die Schöpfungshöhe des Programms in Frage stellen. Der KI selbst Rechte an den Ergebnissen einzuräumen, würde offenlassen, wer über die Rechteeinräumung an Dritte entscheiden soll (Legner 2019): Die KI oder die Schöpfer der KI? Oder auch die Nutzer?

Unvorhersehbarkeit und Intransparenz dürfen nicht dazu führen, dass das Recht den Innovationszyklen der KI hinterherhinkt.

Grundsätzlich erfordert der Investitionsschutz nach dem geltenden Urheberrechtsgesetz eine menschliche Schöpfung. Je weiter sich das Schöpfungsergebnis durch den Einsatz von KI von den Vorgaben der Programmierung entwickelt, umso weniger kann der Schutz dem Urheber als menschlichem Schöpfenden zugerechnet werden (Specht-Riemenschneider 2021). Um die Rechtsunsicherheiten beim Investitionsschutz zu beseitigen, wird für die Softwareerzeugnisse der KI u. a. ein Schutzrecht für Algorithmenerzeugnisse gefordert (Specht-Riemenschneider 2021).

Eigentumsrechtliche Zuordnung der Datensätze

Neben den neu geschaffenen Softwareroutinen ist zu fragen, wer Inhaber der durch die KI geschaffenen Daten sein kann. Grundsätzlich geht die Rechtsprechung davon aus, dass Software eine Sache darstellt, da sie einen binären physikalischen Ladungszustand auf einem Datenträger repräsentiert. Folgt man dieser Auffassung und wendet sie auch auf sonstige Daten an, könnten die Inhaber der Hardware Ausschlussrechte gegenüber Dritten an den Ergebnissen der KI geltend machen. Gegen die dingliche beziehungsweise eigentumsrechtliche Zuordnung von Datensätzen wendet sich der Bundesbeauftragte für Datenschutz und Informationssicherheit mit der Forderung, dass statt „des verdinglichenden Datenbegriffes im Sinne eines ‚Dateneigentums' […] eine Datenwirtschaft den Leitbegriff der Information und damit auch die Wissensperspektive" (BfDI 2021, S. 2) betont werden sollten, damit „die gesellschaftlichen Herausforderungen […] mit Blick auf Daten als öffentliches Gut und die Potenziale von Open Source, Open Data und Stärkung von Demokratiestrukturen besser sichtbar" (ebd.) würden. Teils werden auch Rechte der Skribenten der Daten (also derjenigen, welche faktisch, etwa durch das Festlegen und Befahren einer Navigationsroute im Pkw, die Datenspur ‚schreiben') und der Datenbesitzer als Anhaltspunkt formuliert (Hoeren 2019).

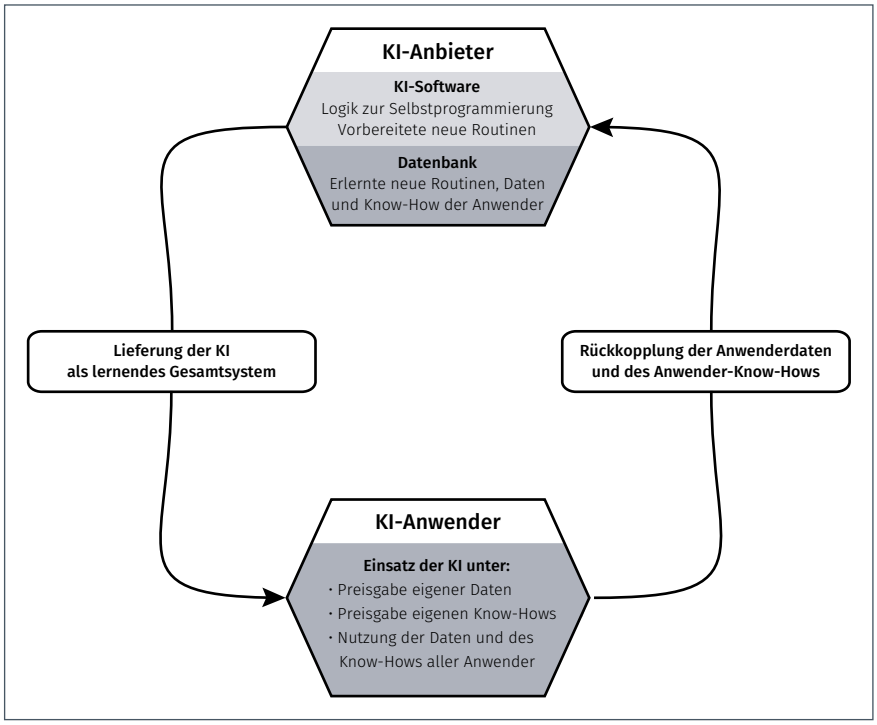


Abb. 1: Schematische KI-Darstellung für KI mit fortlaufender Rückkoppelung der Ergebnisse.
Falls man den Urheber des Programms nicht als Schöpfer des Ergebnisses der KI betrachtet, würde es sich bei den Ergebnissen (einschließlich neuer Algorithmen) nicht mehr um eine ‚menschliche' Schöpfung im Sinne des Urhebergesetzes handeln (Ory und Sorge 2019). Gleiches würde für den nach dem Urheberrecht möglichen Datenbankschutz gelten. In der Konsequenz wären die Ergebnisse dann gemeinfrei und würden der Allgemeinheit zur Verfügung stehen. Die Nichtzuordnung einer kreativen Leistung zu einem Menschen bedeutete damit, dass die Schöpfung von Ergebnissen durch KI nach urheberrechtlichen Maßstäben nicht zu kommerzialisieren wäre. *Quelle: eigene Darstellung*

Allein aus dem Versuch einer eigentumsrechtlichen Zuordnung lassen sich keine adäquaten Lösungen finden, da beispielsweise beim Wechsel des Geschäftsmodells vom Verkaufs eines smarten Produkts (samt KI) zu einer Vermietung der Systembetreiber die Zuordnung des Eigentums bzw. des Besitzes allein bestimmen könnte. Auch der Besitz als Anhaltspunkt ist manipulierbar (etwa durch die Verlegung der Daten in die Cloud), sodass die Skribenteneigenschaft noch am ehesten eine Rechtezuordnung an den Nutzer der KI erlaubt.

Schutz personenbezogener Daten

Datenschutzrechtlich können die Ergebnisse der KI nur relevant sein, wenn sie auf eine natürliche Person bezogen oder beziehbar sind. Ob dies zutrifft, kann nur anhand des Einsatzes der KI beurteilt werden. Liegt ein Personenbezug der auszuwertenden Daten vor, können KI-Einsätze nur auf Basis einer Rechtsgrundlage (u. a. Einwilligung, Vertragserfüllung, berechtigtes Interesse etc.) erfolgen. Bereits 15,9 % der 2019 vom Bundesverband der Personalmanager im März 2019 befragten Unternehmen setzen KI-Anwendungen in der Personalarbeit ein (BPM 2019). Nach den europäischen Maßstäben der DSGVO erfordert ein solcher Einsatz neben einer Rechtsgrundlage die Offenlegung der Scoring-Berechnungen und die Anwendung der Antidiskriminierungsgrundsätze des Allgemeinen Gleichbehandlungsgesetzes. Zu den Risiken im Datenschutzbereich wird auf die ‚Opazität' der KI verwiesen, welche keine transparenten Vorhersagen über das Verhalten der KI zulasse (Europäische Kommission 2020 a, S. 14). Als KI-immanente Risiken gelten Unvorhersehbarkeit und Intransparenz (Meyer 2018), diese dürfen jedoch nicht dazu führen, dass das Recht den Innovationszyklen der KI hinterherhinkt.

Problematisch kann der KI-Einsatz in Arbeitsverhältnissen sein, in welchem KI-Einsatz aufgrund einer Einwilligung – trotz des möglichen Machtungleichgewichts – denkbar ist. Nach einem datenschutz- und grundrechtlich kritisierten Urteil des Landesarbeitsgerichts München[1] kann sogar die Verarbeitung personenbezogener Beschäftigtendaten durch eine KI-Software

1 Beschluss vom 23. 07. 2020–2 Tabakverordnung 126/19.

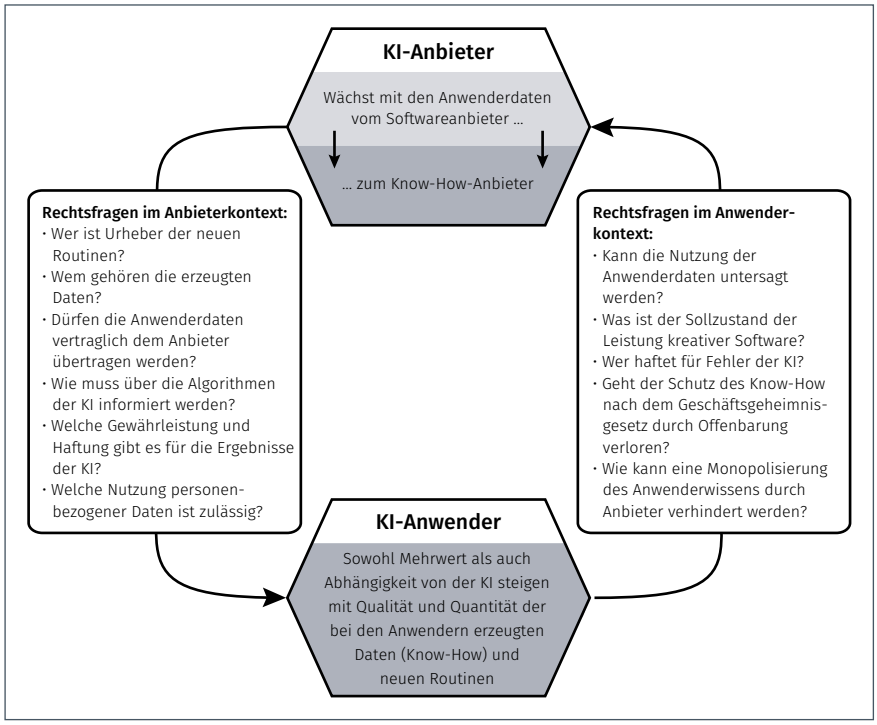


Abb. 2: Formaler Knowhow-Verlust durch Offenbarung des Knowhows an KI. *Quelle: eigene Darstellung*

für präventive Zwecke zulässig sein, wenn es darum geht, Auffälligkeiten des individuellen Arbeitsverhaltens zu erkennen (Wedde 2021).

Eine Einwilligung in die KI-Nutzung und entsprechende Big-Data-Anwendungen wird allerdings in aller Regel an den erforderlichen datenschutzrechtlichen Transparenzvorgaben scheitern. Die Zwecke der durch eine Einwilligung erlaubten Datenverarbeitung nach Art. 6 I a/Art. 7 DSGVO müssen konkret benannt werden. Dies gebietet auch das Prinzip der Zweckbindung nach Art. 5 I b DSGVO, was bei einer echten Big Data-Datenaggregation naturgemäß nicht der Fall sein wird, da es gerade die Idee einer Datenaggregation ist, noch unbekannte spätere Analyse und Netzwerkeffekte zu generieren (Holthausen 2021).

Diese Widersprüche zum Wunsch nach einer Big-Data-Nutzung in Europa lassen sich schwer auflösen. Nachdem sich die deutsche Bundesregierung mit den Eckpunkten einer Datenstrategie bemüht hat, eine Kommerzialisierung der Datenauswertung nicht grundsätzlich abzulehnen, wird diese Strategie vom Bundesbeauftragten für Datenschutz kritisch hinterfragt (BfDI 2021, S. 2). Auch die Idee, dem Konflikt mit dem Datenschutz durch eine Anonymisierung personenbezogener Daten zu entgehen, welche die Daten dem Anwendungsbereich der DSGVO entziehen würde, ist umstritten. Teilweise wird nämlich vertreten, dass die Anonymisierung selbst ebenfalls einer datenschutzrechtlichen Rechtsgrundlage bedarf (Hornung und Wagner 2020). Datenschutz kann jedenfalls dann Big-Data-Anwendungen entgegenstehen, soweit keine vollständige Anonymisierung der Daten vorgenommen werden kann (Roßnagel 2013). Selbst wenn eine Anonymisierung gelingt, sind Beeinträchtigungen der Grundrechtsausübung nicht ausgeschlossen. Roßnagel (2013, S. 566) weist für bestimmte Einsatzszenarien darauf hin, dass KI-Analysen etwa dazu führen können, dass „durch eine Big-Data-Analyse bekannt wird, dass Angehörige einer politischen Gruppe sich überdurchschnittlich oft an einem bestimmten Ort aufhalten, in einem bestimmten Geschäft einkaufen und ÖPNV fahren“. Dies könne wiederum dazu führen, dass „bei vielen, die vermeiden wollen, dieser Gruppe zugerechnet zu werden, […] dass sie den Ort, das Geschäft und vielleicht sogar das Verkehrsmittel meiden, um nicht in einen ihnen unangenehmen Verdacht zu geraten“ (ebd.). Damit verstärkten solche Analysen „die Normativität der Normalität“ (ebd.) und reduzierten die für die Demokratie notwendige ‚Soziodiversität‘.

Das Datenschutzrecht bietet somit keine befriedigende Auskunft über die Grenzen der Nutzungsmöglichkeit der KI, da der von dieser generierte Datensatz – durch Anonymisierung den Regelungen der Datenschutzgrundverordnung und weiterer in Verabschiedung befindlicher E-Privacy-Regelungen (Council of the European Union 2021) – der Kontrolle der Betroffenen entzogen werden kann.

Versuche der Monopolisierung der KI-Ergebnisse

Vertragsrechtlich sind in der Praxis Versuche zu verzeichnen, die Nutzern der KI zu verpflichten, die KI-Ergebnisse den Herstellenden zu übertragen, damit diese dann Zugriff auf die Ergebnisse erhalten und entsprechende Big Data-Anwendungen erstellen, um die Daten anderweitig nutzen zu können. Grenzen dieser Rückübertragungsversuche finden sich im Recht der Allgemeinen Geschäftsbedingungen und des Kartellrechts.

De lege lata ist mithin festzuhalten, dass die KI-Ergebnisse nicht ohne Weiteres den Herstellenden zustehen. Da jedoch die KI vertragliche Rahmenbedingungen voraussetzt, wird in aller Regel auch die Rückkopplung der KI-Ergebnisse den Herstellenden zu Gute kommen, siehe Abb. 2. Gesetzliche Haftungsregelungen wiederum werden den multikausalen Zusammenhängen bei KI-Ergebnissen und der Zersplitterung der Haftungszuordnung angesichts der Beweisnöte der Anwendern nicht gerecht.

Daher bedarf es einer Regelung der Datensouveränität der Betroffenen, die die Wissensagglomeration der Betreiber der KI-Systeme angemessen einbezieht.

Gesellschaftliche Folgen des KI-Einsatzes und Lösungsansätze

Angesichts der unübersichtlichen und zu komplexen rechtlichen Situation (Kühling und Sackmann 2020) besteht weiterer Handlungsbedarf zum Umgang mit der KI, eine lediglich beobachtende Haltung wird nicht genügen.

De lege ferenda werden Versuche unternommen, in einzelnen marktrelevanten Bereichen die Offenlegung von Algorithmen vorzugeben, welche beispielsweise die Funktionsweise digitaler Marktplätze und die Information über KI-Funktionsweisen betreffen:

Deutsche Gesetzgebung

Im Gesetzentwurf der Bundesregierung „zur besseren Durchsetzung und Modernisierung der Verbraucherschutzvorschriften der Union“ (Deutscher Bundestag 2021) werden in Artikel 246 d § 1 BGB-neu allgemeine Informationspflichten für Betreiber von Online-Marktplätzen festgelegt, etwa über die „Hauptparameter zur Festlegung des Rankings und die relative Gewichtung der Hauptparameter zur Festlegung des Rankings im Vergleich zu anderen Parametern“ zu informieren. Die Reform des Netzwerkdurchsetzungsgesetzes (NetzDG), welches sich mit der Bekämpfung von ‚Hate Speech‘ auf sozialen Plattformen befasst, greift die KI-Problematik auf, indem eine in § 2 Absatz 2 NetzDG festgelegte neue Informationspflicht vorsieht, dass über „Art, Grundzüge der Funktionsweise und Reichweite von gegebenenfalls eingesetzten Verfahren zur automatisierten Erkennung von Inhalten“ Auskunft gegeben werden muss.

Weiterhin wird in einer Neuregelung des Kartellrechts (§ 19 Abs. 2 Nr. 1 GWB) vorgesehen, dass unter bestimmten Umständen den Zugang zu Daten und gegebenenfalls auch KI-Datenbanken eröffnet werden muss (Huerkamp und Nuys 2021). Weiterhin wird im Betriebsrätemodernisierungsgesetz zur Stärkung der Stellung des Betriebsrats bei der Beurteilung von KI-Einsätzen die Hinzuziehung einer sachverständigen Person ermöglicht.

Europäische Ansätze

Umfassender als diese zersplitterten nationalen Ansätze fallen, wie im Folgenden dargestellt, europäische Regelungen zur KI-Kontrolle aus:

Artificial Intelligence Act

In diesem Verordnungsvorschlag (COM/2021/206 final) sollen unter anderem KI-Systeme verboten werden (Art. 5 des Verordnungsentwurfs), die Techniken der unterschwelligen Beeinflussung außerhalb des Bewusstseins einer Person einsetzen oder eine Schwäche beziehungsweise Schutzbedürftigkeit einer bestimmten Gruppe von Personen ausnutzen sowie Systeme zur Klassifizierung der Vertrauenswürdigkeit natürlicher Personen und biometrische Echtzeit-Fernidentifizierungssysteme. Daneben sind Regelungen zur transparenten Information, zum Risikomanagement, zu Daten und Daten-Governance, zur menschlichen Aufsicht und zu weiteren Sicherheitsmaßnahmen bei ‚Hochrisiko‘-KI-Systemen vorgesehen.

Data Governance Act, Data Act

Mit dem Data Governance Act (Europäische Kommission 2020 b) soll generell die Bereitstellung von Daten geregelt werden. Dies gilt für Daten des öffentlichen Sektors zur Weiterverwendung in Fällen, in denen diese Daten den Rechten anderer unterliegen, die gemeinsame Datennutzung durch Unternehmen gegen Entgelt in jedweder Form, die Ermöglichung der Nutzung personenbezogener Daten mithilfe eines „Mittlers für die gemeinsame Nutzung personenbezogener Daten“ (Europäische Kommission 2020 b), der Einzelpersonen bei der Ausübung ihrer Rechte gemäß der Datenschutz-Grundverordnung (DSGVO) unterstützen soll, sowie die Ermöglichung der Nutzung von Daten aus altruistischen Gründen. Er soll nicht darauf abzielen, „wesentliche Rechte auf den Zugang zu Daten und deren Nutzung zu gewähren, zu ändern oder zu beseitigen“ (Europäische Kommission 2020 b). Es ist geplant, diese Art von Maßnahmen in einen möglichen Rechtsakt über Daten (2021) aufzunehmen (European Parliament 2021).

In Erwägungsgrund 6 des Data Governance Act wird darauf verwiesen, dass es für Datenbanken, die personenbezogene Daten enthalten, Techniken gibt, „die datenschutzfreundliche Analysen ermöglichen, zum Beispiel Anonymisierung, Pseudonymisierung, differentielle Privatsphäre, Generalisierung oder Datenunterdrückung und Randomisierung“ (Europäische Kommission 2020 b). Mithilfe dieser soll eine sichere Weiterverwendung personenbezogener Daten und vertraulicher Geschäftsdaten für Forschung, Innovation und statistische Zwecke gewährleistet werden können. Schließlich heißt es, dass die Verarbeitung personenbezogener Daten stets auf einem der Rechtsgründe beruhen solle, die in Artikel 6 DSGVO aufgeführt sind.

Die Abgrenzung der DSGVO zum Data Governance Act bleibt leider unklar (Graef et al. 2020), da nicht eindeutig geklärt ist, ob der Data Governance Act auch pseudonyme Daten erfasst. Die neue Gesetzgebung „wird die im Rahmen der Richtlinie über offene Daten noch zu erlassenden Vorschriften über hochwertige Datensätze ergänzen, die EU-weit den kostenlosen Zugang zu bestimmten Datensätzen in maschinenlesbarem Format und über genormte Anwendungsprogrammierschnittstellen […] regeln sollen. […] Für 2021 sind weitere konkrete Vorschläge für besondere Datenräume geplant, beispielsweise für einen europäischen Gesundheitsdatenraum und einen Datenraum für den Grünen Deal. All dies wird ergänzt durch einen neuen Rechtsakt über Daten, der den Bürgerinnen und Bürgern sowie den Behörden einen besseren Zugang zu Daten aus dem sogenannten Internet of Things der Industrie und zu Massendaten im Besitz von Unternehmen sichern und eine bessere Kontrolle über solche Daten verschaffen soll, um eine gerechtere Wirtschaft aufzubauen und Vorteile für die Gesellschaft insgesamt zu erzielen.” (Europäische Kommission 2020 c, 4) Soweit diese Regelungen parallel nebeneinander bestehen bleiben, wird der Wunsch nach einer transparenten Regelung jedoch unerhört bleiben.

Vorschläge zur vertraglichen Regulierung, zu Control by Design und Technikgestaltung

Politisch wünschbare Regulierungsansätze – und insbesondere der geplante europäische ‚Data Act' sollten daher Grenzen der Auswertbarkeit der KI-Ergebnisse durch die Systembetreiber festlegen, einen transparenten Einsatz der KI insbesondere in sensiblen persönlichkeitsrechtlichen Bereichen der Analyse (Medizin und Arbeitswelt) festlegen und arbeitsweltliche Grenzen des Ersatzes menschlicher Arbeitsleistung durch KI zumindest anstreben.

Denkbar wäre eine übergreifende Regulierung auf europäischer Ebene, welche technikorientiert, ähnlich wie die aktuell (erneut) in Verabschiedung befindliche E-Privacy-Verordnung, und szenariobasiert die Risiken der KI in den Blick nimmt und die erwähnten rechtliche Lücken schließt. Zugleich muss sie jedoch das Verhältnis zur Datenschutzgrundverordnung klarer als im Data Governance Act regeln und auch eine Einwilligung in noch nicht bekannte Auswertungsarten und Daten mit verschiedenen Stufen des Personenbezugs in bestimmten Fällen zulassen, wenn KI und Big Data ernsthaft betrieben werden sollen.

Hierzu kämen zwei Ansätze in Betracht:

Klauselregelungen für Datenübertragungen

Da die dargestellten gesetzlichen Regulierungsansätze völlig heterogene Fälle der Information der Nutzer über den KI-Einsatz (abhängig von der Art der Daten) vorsehen, sollte eine Rahmenregelung für zulässige Datenauswertungen sowohl für personenbezogene als auch für anonymisierte Daten vorgesehen werden (Schur 2020). Diese Klauselregelungen könnten sich an den AGB-Klauselregelungen der §§ 308, 309 BGB oder den sog. ‚grauen' und ‚schwarzen' Klauseln der EU-Gruppenfreistellungsverordnungen orientieren. So könnten – orientiert an den bekannten Technikfolgen der KI – die Grenzen einwilligungsfähiger personenbezogener Datenauswertungen und nicht personenbezogener Datenauswertungen unterschieden werden. Dies würde zugleich einen klaren Rahmen für die Ersteller und Betreiber von KI-Systemen darstellen, die Rechtssicherheit über die wirksamen Vereinbarungen mit den Nutzern erhalten. Denkbar wäre hier auch eine Beteiligung der Nutzer an der Wertschöpfung abhängig vom Umfang der Einwilligung in die Datenauswertung.

Control by Design

Analog zu den Regelungen der DSGVO für Privacy by Design könnte auch für nicht personenbezogene Daten – begrenzt durch die und angebunden an die oben erwähnten Klauselregelungen – eine Control by Design-Regelung eingeführt werden (der Begriff ‚Privacy' träfe hier nicht zu, da auch anonyme Auswertungen von der Datensouveränität umfasst wären).

So wie das im Entwurf vorliegende Gesetz über den Datenschutz und den Schutz der Privatsphäre in der Telekommunikation und bei Telemedien (TTDSG) in seinen §§ 25, 26 den Einsatz von ‚Personal Information Management Services' regelt und eine transparente Einwilligung der Freigabe von Endnutzerdaten (etwa bei Cookies) ermöglicht, könnte auch im Data Act die im TTDSG vorgesehene Einschaltung von „Anerkannten Diensten zur Einwilligungsverwaltung" (§ 26 TTDSG) erfolgen.

In der konkreten Ausprägung der Einwilligung könnten auch banale Fragen wie die Haftung für Ergebnisse der KI-Aktivitäten gegenüber den beteiligten Akteuren (Hacker 2020; Zech 2019) oder die Übertragung der Scoring-Transparenz der DSGVO für nicht-personenbezogene KI-Daten berücksichtigt werden.

Die EU versucht in ihren Regulierungsansätzen, der Marktmonopolisierung im KI-Umfeld und der Monetarisierung der Nutzerdaten entgegenzuwirken, um auch altruistische Datenverwendungen zu ermöglichen und bei der Anwendung der KI die Grundrechte der Nutzer zu beachten. Eine entsprechende Umsetzung in der Praxis müsste sich jedoch am Kriterium der Umsetzbarkeit messen lassen, wenn sie erfolgreich sein soll.

Mit einer solchen – über den Data Governance Act hinausgehenden – Regelung des Control by Design wäre es denkbar, der Datensouveränität der Nutzer gerecht zu werden und zugleich eine sichere Rechtsgrundlage für die Betreiber der Systeme zu schaffen.

Angabe von Finanzierungsquellen

Der vorliegende Forschungsartikel hat keine Förderung erhalten.

Literatur

BfDI – Der Bundesbeauftragte für den Datenschutz und die Informationsfreiheit (2021): Stellungnahme des Bundesbeauftragten für den Datenschutz und die Informationsfreiheit zur öffentlichen Anhörung des Ausschusses Digitale Agenda am 24. Februar 2021. Online verfügbar unter https://www.bfdi.bund.de/SharedDocs/Downloads/DE/DokumenteBfDI/Stellungnahmen/2021/StgN_StgN_Anh%C3%B6rung-Datenstrategie-Bundesregierung.pdf;jsessionid=7D1B64827D3B8DA4D33C87024D824D47.intranet242?__blob=publicationFile&v=4, zuletzt geprüft am 21. 10. 2021, S. 1.

BPM – Bundesverband der Personalmanager (2019): Künstliche Intelligenz in der Personalarbeit. Online verfügbar unter https://www.bpm.de/sites/default/files/20190429_auswertung_bpm-pressemitteilung_final_0.pdf, zuletzt aufgerufen am 18. 10. 2021.

Council of the European Union (2021): Proposal for a regulation of the European Parliament and of the council concerning the respect for private life and the protection of personal data in electronic communications and repealing directive 2002/58/EC (regulation on privacy and electronic communications). Online verfügbar unter https://data.consilium.europa.eu/doc/document/ST-6087-2021-INIT/en/pdf, zuletzt geprüft am 21. 10. 2021.

Datenethikkommission (2019): Gutachten der Datenethikkommission der Bundesregierung, Potsdam, S. 224. Online verfügbar unter https://datenethikkommission.de/wp-content/uploads/191015_DEK_Gutachten_screen.pdf, zuletzt geprüft am 26. 09. 2021.

Deutscher Bundestag (2021): Drucksache 19/27 655. Berlin: H. Heenemann. Online verfügbar unter https://dserver.bundestag.de/btd/19/276/1927655.pdf, zuletzt geprüft am 21. 10. 2021.

Europäische Kommission (2020 a): Weißbuch zur Künstlichen Intelligenz. Ein europäisches Konzept für Exzellenz und Vertrauen. Online verfügbar unter

https://ec.europa.eu/info/sites/default/files/commission-white-paper-artificial-intelligence-feb2020_de.pdf, zuletzt geprüft am 10. 11. 2021.

Europäische Kommission (2020 b): Vorschlag für eine Verordnung des Europäischen Parlaments und des Rates über europäische Daten-Governance (Daten-Governance-Gesetz). Online verfügbar unter https://eur-lex.europa.eu/legal-content/DE/TXT/HTML/?uri=CELEX:52020PC0767&from=EN, zuletzt geprüft am 08. 11. 2021.

Europäische Kommission (2020 c): Verordnung über Daten-Governance – Fragen und Antworten. Online verfügbar unter https://ec.europa.eu/commission/presscorner/api/files/document/print/de/qanda_20_2103/QANDA_20_2103_DE.pdf, zuletzt geprüft am 25. 11. 2021.

European Parliament (2021): Legislative Train 09.2021. 2 a Europe fit for the digital age. Online verfügbar unter https://www.europarl.europa.eu/legislative-train/api/stages/report/current/theme/a-europe-fit-for-the-digital-age/file/data-act, zuletzt geprüft am 21. 10. 2021.

Graef, Inge; Husovec, Martin; van den Boom, Jasper (2020): Spill-overs in data governance. Uncovering the uneasy relationship between the GDPR's right to data portability and EU sector-specific data access regimes. In: Journal of European Consumer and Market Law 9 (1), S. 3–16. https://doi.org/10.2139/ssrn.3369509

Hacker, Philipp (2020): Europäische und nationale Regulierung von Künstlicher Intelligenz. In: Neue Juristische Wochenschrift 73 (30), S. 2142–2147. Online verfügbar unter https://beck-online.beck.de/Dokument?vpath=bibdata%2Fzeits%2Fnjw%2F2020%2Fcont%2Fnjw.2020.2142.1.htm&pos=1&hlwords=on, zuletzt geprüft am 21. 10. 2021.

Herberger, Maximilian (2018): „Künstliche Intelligenz" und Recht. In: Neue Juristische Wochenschrift 71 (39), S. 2825–2829. Online verfügbar unter https://beck-online.beck.de/Dokument?vpath=bibdata%2Fzeits%2Fnjw%2F2018%2Fcont%2Fnjw.2018.2825.1.htm&pos=2&hlwords=on, zuletzt geprüft am 21. 10. 2021.

Hoeren, Thomas (2019): Datenbesitz statt Dateneigentum. In: Multimedia und Recht 22 (1), S. 5–8. Online verfügbar unter https://beck-online.beck.de/Dokument?vpath=bibdata%2Fzeits%2Fmmr%2F2019%2Fcont%2Fmmr.2019.5.1.htm&pos=1&hlwords=on, zuletzt geprüft am 21. 10. 2021.

Holthausen, Joachim (2021): Big data, people analytics, KI und Gestaltung von Betriebsvereinbarungen. Grund-, arbeits- und datenschutzrechtliche An- und Herausforderungen. In: Recht der Arbeit 74 (1), S. 19–32. Online verfügbar unter https://beck-online.beck.de/Dokument?vpath=bibdata%2Fzeits%2Frda%2F2021%2Fcont%2Frda.2021.19.1.htm&pos=3, zuletzt geprüft am 21. 10. 2021.

Hornung, Gerrit; Wagner, Bernd (2020): Anonymisierung als datenschutzrelevante Verarbeitung? In: Zeitschrift für Datenschutz 10 (5), S. 223–228. Online verfügbar unter https://beck-online.beck.de/Dokument?vpath=bibdata%2Fzeits%2Fzd%2F2020%2Fcont%2Fzd.2020.223.1.htm&anchor=Y-300-Z-ZD-B-2020-S-223-N-1, zuletzt geprüft am 21. 10. 2021.

Huerkamp, Florian; Nuys, Marcel (2021): Datenzugang nach § 19 Abs. 2 Nr. 4 GWB n. F. – Geglückte „Klarstellung"? In: Neue Zeitschrift für Kartellrecht 9 (7), S. 327–329. Online verfügbar unter https://beck-online.beck.de/Dokument?vpath=bibdata%2Fzeits%2Fnzkart%2F2021%2Fcont%2Fnzkart.2021.327.1.htm&pos=7, zuletzt geprüft am 21. 10. 2021.

Kühling, Jürgen; Sackmann, Florian (2020): Irrweg Dateneigentum. In Zeitschrift für Datenschutz 10 (1), S. 24–30. Online verfügbar unter https://beck-online.beck.de/Dokument?vpath=bibdata%2Fzeits%2Fzd%2F2020%2Fcont%2Fzd.2020.24.1.htm&pos=16, zuletzt geprüft am 21. 10. 2021.

Legner, Sarah (2019): Erzeugnisse Künstlicher Intelligenz im Urheberrecht. In: Zeitschrift für Urheber und Medienrecht 63 (11), S. 807–812. Online verfügbar unter https://www.researchgate.net/profile/Sarah-Legner-2/publication/343510297_Erzeugnisse_Kunstlicher_Intelligenz_im_Urheberrecht/links/5fb4d44fa6fdcc9ae05ef8c8/Erzeugnisse-Kuenstlicher-Intelligenz-im-Urheberrecht.pdf, zuletzt geprüft am 21. 10. 2021.

Meyer, Stephan (2018): Künstliche Intelligenz und die Rolle des Rechts für Innovation. In: Zeitschrift für Rechtspolitik 51 (8), S. 233–237. Online verfügbar unter https://beck-online.beck.de/?vpath=bibdata%2fzeits%2fZRP%2fcont%2fZRP%2ehtm, zuletzt geprüft am 21. 10. 2021.

Ory, Stephan; Sorge, Christoph (2019): Schöpfung durch Künstliche Intelligenz? In: Neue Juristische Wochenschrift 72 (11), S. 710–712. Online verfügbar unter https://beck-online.beck.de/?vpath=bibdata%2Fzeits%2FNJW%2F2019%2Fcont%2FNJW%2E2019%2E710%2E1%2Ehtm, zuletzt geprüft am 21. 10. 2021.

Rauer, Nils; Ettig, Diana (2021): Update Cookies 2020. Aktuelle Rechtslage und Entwicklungen. In: Zeitschrift für Datenschutz 11 (1), S. 18–24. Online verfügbar unter https://beck-online.beck.de/Dokument?vpath=bibdata%2Fzeits%2Fzd%2F2021%2Fcont%2Fzd.2021.18.1.htm&pos=2&hlwords=on, zuletzt geprüft am 21. 10. 2021.

Roßnagel, Alexander (2013): Big Data – Small Privacy? Konzeptionelle Herausforderungen für das Datenschutzrecht. In: Zeitschrift für Datenschutz 3 (11), S. 562–567. Online verfügbar unter https://beck-online.beck.de/Dokument?vpath=bibdata%2Fzeits%2Fzd%2F2013%2Fcont%2Fzd.2013.562.1.htm&anchor=Y-300-Z-ZD-B-2013-S-562-N-1, zuletzt geprüft am 21. 10. 2021.

Schur, Nico (2020): Die Lizenzierung von Daten. Einordnung, Grenzen und Möglichkeiten von vertraglichen Zugangs- und Datennutzungsrechten in der digitalen Ökonomie. In: Jeanette Hofmann; Ingolf Pernice; Thomas Schildhauer; Wolfgang Schulz (Hg.): Internet und Gesellschaft. Schriften des Alexander von Humboldt Institut für Internet und Gesellschaft: Tübingen Mohr Siebeck.

Specht-Riemenschneider, Louisa (2021): Urheberrechtlicher Schutz für Algorithmenerzeugnisse? Phasenmodell de lege lata, Investitionsschutz de lege ferenda? In: Wettbewerb in Recht und Praxis 18 (3), S. 273–275.

Statista Research Department (2020): Umfrage zum Umgang mit Cookie-Hinweisen in Deutschland 2020. Online verfügbar unter https://de.statista.com/statistik/daten/studie/1121071/umfrage/umgang-mit-cookie-hinweisen-in-deutschland/, zuletzt geprüft am 13. 10. 2021.

Wedde, Peter (2021: Anmerkung – Streit um Einigungsstellenspruch zur Einführung eines IT-Sicherheitssystems. Anlasslose präventive Verarbeitung von Beschäftigtendaten durch KI-Software zulässig, LArbG München v. 23. 07. 2020–2 TaBV 126/19, jurisPR-ArbR 17/2021 Anm. 6. In: Juris – Das Rechtsportal.

Zech, Herbert (2019): Künstliche Intelligenz und Haftungsfragen. In: Zeitschrift für die gesamte Privatrechtswissenschaft 5 (2), S. 198–219.

PROF. DR. THOMAS WILMER
ist seit 2002 Professor für Informationsrecht an der Hochschule Darmstadt und leitet dort das Institut für Informationsrecht. Er befasst sich mit Fragen des Lizenz- und Datenschutzrechts und verwandter Bereiche des Informationsrechts.