MICHAEL REDER,
CHRISTOPHER KOSKA (HG.)

KÜNSTLICHE INTELLIGENZ UND ETHISCHE VERANTWORTUNG

[transcript] Edition Moderne Postmoderne

Michael Reder, Christopher Koska (Hg.)
Künstliche Intelligenz und ethische Verantwortung

Edition Moderne Postmoderne

Editorial

Die **Edition Moderne Postmoderne** präsentiert die moderne Philosophie in zweierlei Hinsicht: zum einen als philosophiehistorische Epoche, die mit dem Ende des Hegel'schen Systems einsetzt und als Teil des Hegel'schen Erbes den ersten philosophischen Begriff der Moderne mit sich führt; zum anderen als Form des Philosophierens, in dem die Modernität der Zeit selbst immer stärker in den Vordergrund der philosophischen Reflexion in ihren verschiedenen Varianten rückt – bis hin zu ihrer »postmodernen« Überbietung.

Michael Reder (Dr. phil.) ist Professor für Praktische Philosophie und Vizepräsident für Forschung an der Hochschule für Philosophie München. Er ist Konsortialführer des vom *bidt* finanzierten Forschungsverbundes KAIMo (Kann ein Algorithmus im Konflikt moralisch kalkulieren?) und Mitglied des Direktoriums des gemeinsamen Zentrums für verantwortliche KI (CReAITech) der Technischen Universität München, der Universität Augsburg und der Hochschule für Philosophie München.
Christopher Koska (Dr. phil.) ist wissenschaftlicher Mitarbeiter am Lehrstuhl für Praktische Philosophie der Hochschule für Philosophie München und Partner bei der Unternehmensberatung dimension2 GmbH. Er ist Projektkoordinator des vom *bidt* finanzierten Forschungsprojekts KAIMo (Kann ein Algorithmus im Konflikt moralisch kalkulieren?) und Postdoc am Center for Responsible AI Technologies (CReAITech). Sein Forschungs- und Arbeitsschwerpunkt ist das Themenfeld der Daten- und Algorithmenethik sowie deren Umsetzung im Kontext der Corporate Digital Responsibility (CDR).

Michael Reder, Christopher Koska (Hg.)

Künstliche Intelligenz und ethische Verantwortung

[transcript]

Wir danken dem Bayerischen Forschungsinstitut für Digitale Transformation (bidt) für die Unterstützung und Finanzierung des Forschungsprojekts KAIMo (Kann ein Algorithmus im Konflikt moralisch kalkulieren?) sowie des vorliegenden Bandes.

Bibliografische Information der Deutschen Nationalbibliothek
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über https://dnb.dnb.de/ abrufbar.

Erschienen 2024 im transcript Verlag, Bielefeld

Umschlaggestaltung: Kordula Röckenhaus, Bielefeld
Lektorat: Daniel Seltier und Kilian Porth
Korrektorat: Sophie Wax
Druck: Majuskel Medienproduktion GmbH, Wetzlar
https://doi.org/10.14361/9783839469057
Print-ISBN: 978-3-8376-6905-3
PDF-ISBN: 978-3-8394-6905-7
Buchreihen-ISSN: 2702-900X
Buchreihen-eISSN: 2702-9018

Gedruckt auf alterungsbeständigem Papier mit chlorfrei gebleichtem Zellstoff.

Inhalt

Über das Verhältnis von Ethik und Algorithmen
Ein Problemaufriss

Michael Reder, Nicholas Müller, Robert Lehmann

Gegenwärtig wird intensiv darüber diskutiert, ob künstliche Intelligenz eine sehr spezifische menschliche Eigenschaft übernehmen kann, und zwar die, moralisch zu handeln (Floridi und Sanders 2004; Misselhorn 2018; Brieger 2018; Weber 2018). Hintergrund dieser Diskussion sind technische Entwicklungen, die der Maschine scheinbar mehr und mehr so etwas wie autonomes Handeln ermöglichen. Maschinen haben im Zuge dessen nicht nur automatisierte Prozesse übernommen, sondern durch die ihnen spezifische Trainings- und Lernfähigkeit auch neue Aktionsmöglichkeiten erworben, die handlungsanalog zu sein scheinen. Mit diesen Möglichkeiten zu eigenständigem Handeln wird nun die Frage relevant, ob die Maschine auch etwas tun kann, was oftmals nur dem Menschen zugeschrieben wird: und zwar moralisch zu handeln. Im Kern der Diskussionen steht die Frage, ob Maschinen moralisch kalkulieren, entscheiden und handeln können und ob ihnen deshalb auch Verantwortung für ihr Handeln zu geschrieben werden kann (Dignum 2018).

Katharina Zweig (2018) hat in diesem Zusammenhang auf einige grundlegende Probleme hingewiesen. Denn die Informatik braucht eine formalisierte Definition von normativen Begriffen. Erst dann können diese in einen digitalen Code übersetzt werden. Die philosophische Frage nach der Normativität ist jedoch selten so binär, wie die Informatik sie gerne hätte. Normative Konflikte sind oft sehr kompliziert, auch weil sie sich zeitlich verändern und deswegen nur selten in einer eindeutigen Heuristik konzeptualisiert werden können. Dies gilt auch für Normen selbst. Viele Normen sind oft deshalb so formal, weil nur so ihre Allgemeingültigkeit begründet werden kann. Wenn diese Normen material gefüllt werden, ist diese Eindeutigkeit allerdings begrenzt. Mit Blick auf verschiedene soziale, kulturelle oder zeitliche Kontexte einerseits und mit Blick auf den Einzelfall andererseits werden Normen oft unterschiedlich ma-

terial gefüllt. Diese Kontextualität und inhärente Dialektik von Normativität widersprechen der binären Logik der digitalen Technologie.

Der vorliegende Band will diese Frage nach dem Verhältnis von KI und ethischer Verantwortung von unterschiedlichen Seiten aus – teils anhand von Fallbeispielen, teils in systematischer Hinsicht – analysieren und kritisch diskutieren. Es geht darum, Chancen und Risiken beim Umgang mit KI-Technologien auszuloten und nach ethisch verantwortlichen Formen des Umgangs, v.a. in Konfliktsituationen, zu suchen.

Dem Band liegen die Arbeiten eines interdisziplinären Forschungsverbundes zugrunde.[1] In diesem geht es um ein Feld institutionellen Handelns angesichts von Konflikten mit massiver Reichweite (Gutwald et al. 2021). Dabei handelt es sich um die Bewertung von Fallakten durch Jugendämter im Hinblick auf Risiken der Kindeswohlgefährdung. Prinzipiell sieht die grundgesetzliche Regelung in Deutschland vor, dass es das Recht und die Pflicht der Eltern ist, ihre Kinder zu erziehen und sich um ihr Wohlergehen zu kümmern. Die Intervention durch staatliche Stellen ist nur dann vorgesehen, wenn eine Gefährdung des Kindeswohls vorliegt, die die Eltern nicht abwenden können oder wollen. Erst dann wird das Wächteramt des Staates relevant und in Verantwortungsgemeinschaft mit dem Familiengericht greift das Jugendamt als öffentlicher Träger der Jugendhilfe in die Familie ein. Die Jugendämter schlagen im begründeten Krisenfall verschiedene Maßnahmen zum Schutz des Kindes vor: von Beratungsdiensten bis zur Inobhutnahme des Kindes. Dazu wurde durch den Bundesgerichtshof konkretisiert, dass eine Kindeswohlgefährdung vorliegt, wenn »eine gegenwärtige, in einem solchen Maße vorhandene Gefahr [festgestellt werden kann], dass sich bei der weiteren Entwicklung eine erhebliche Schädigung mit ziemlicher Sicherheit voraussehen lässt.«(BGH FamRZ 1956: 350).

Der Forschungsverbund untersucht, ob und inwiefern normative Kriterien, die das Handeln von Jugendämtern leiten, in Algorithmen übersetzt werden können und ob digitale Tools das institutionelle Handeln unterstützen können (Gutwald und Reder 2023). Dies ist einerseits angesichts der Komplexität der Situation, des betroffenen Rechtskonflikts und der Reichweite

1 Dabei handelt es sich um den vom bidt finanzierten Forschungsverbund *Kann ein Algorithmus im Konflikt moralisch kalkulieren*, der von 2021–2023 an der Schnittstelle von Philosophie, Informatik und Sozialer Arbeit angesiedelt ist. Das Projekt ist eine Kooperation der Hochschule für Philosophie in München, der Technischen Hochschule Würzburg-Schweinfurt und der Technischen Hochschule Nürnberg Georg Simon Ohm.

der Entscheidung nicht unproblematisch. Es erscheint aber andererseits angesichts eben dieser Tragweite des Konflikts und begrenzter personeller und Zeitressourcen in den Institutionen auch sinnvoll zu fragen, ob digitale Systeme eine Hilfe sein können, um die Entscheidungen mit Bezug auf gesellschaftliche Normen transparent und fundiert zu treffen.

In einer Welt, in der künstliche Intelligenz und Algorithmen immer mehr Entscheidungen übernehmen, stellt sich mit Blick auf solche Beispielfelder und angesichts dieser Eigenart ethischer Urteilsfindung die Frage, inwieweit diese Systeme moralische und ethische Aspekte berücksichtigen können. Eine zentrale Frage ist, ob und wenn ja wie Algorithmen entwickelt werden könnten, die moralisch kalkulieren und in komplexen ethischen Situationen Entscheidungen treffen können. Die folgende Einleitung stellt einen Problemaufriss zu diesem Themenfeld vor dem Hintergrund der Forschungen des genannten Verbundes dar.

Zur Kalkulationsfähigkeit von Algorithmen im Kontext ethischer Entscheidungen

Aus technischer Perspektive gilt es als erstes zu betonen, dass innerhalb des Themenfeldes der KI zwischen den Basistechnologien und Anwendungen zu unterscheiden ist. Christen et al. (2020) postulieren in diesem Zusammenhang, dass vier Basisfunktionen, und zwar Mustererkennung, Klassifikation, Prognose und Synthese aus unterschiedlichen Input-Daten, z.B. Text, Bild oder Ton, einen Anwendungskontext erzeugen. Muster werden in Texten oder Bildern erkannt, Geräusche durch Klassifizierung als Vogelgeräusche präzisiert (Xie et al. 2023), das Wetter durch Datenaggregation vorhergesagt oder neue Musik durch Synthese generiert (Rutherford 2023). Entsprechend weisen Christen et al. (2020) darauf hin, dass durch KI-Technologien Systeme spezifiziert werden können, die mittels Berechnungen Artefakte generieren können. Um diese generierten Daten zu klassifizieren oder solche Systeme für die Gesellschaft nutzbar zu machen, wird jedoch weiterhin menschliche Expertise benötigt. Im Berufsfeld *Clickwork* werden zum Beispiel die Trainingsdaten generiert (Beuth et al. 2023). Im nächsten Schritt wird eine der Basisfunktionen genutzt, um einen neuen Output zu generieren, der wiederum von Menschen genutzt wird, z.B. beim autonomen Fahren. Dieser kann wiederum Ausgangspunkt für weitere menschliche Entscheidungen sein.

Gemeinsam ist diesen Anwendungen, dass sie Daten aggregiert auswerten und eine Schlussfolgerung aus diesen ziehen. Je nach Anwendungsszenario enthalten sowohl die Eingangsdaten als auch die Ergebnisdarstellung Aspekte, die aus ethisch-moralischer Sicht einer besonderen Prüfung bedürfen. Insbesondere dann, wenn diese bereits bei der rechnerbasierten Aggregation der Daten berücksichtigt werden müssen. Denn die Integration von moralischen Aspekten in statistischen Berechnungen ist eine komplexe Herausforderung. Um normative Kriterien in Algorithmen zu übersetzen, müssen sie zunächst quantifiziert werden. Ein Ansatz zur Quantifizierung von moralischen Aspekten besteht darin, ethische Prinzipien in messbare Indikatoren oder Kriterien zu übersetzen und diese in statistische Modelle einzubeziehen. Dabei ist zu berücksichtigen, dass derartige Modelle nur eine Annäherung an die moralische Abwägung darstellen und nicht alle ethischen Nuancen erfassen können. Diese Einschränkungen finden sich auch in Überblicksstudien zur Anwendung von KI-basierten Assistenz- oder Entscheidungssystemen.

Der Bericht von *AlgorithmWatch* (Matzat et al. 2019) leistet an dieser Stelle einen wesentlichen Beitrag, da die dort beschriebene Datenbank eine Übersicht zu vorrangig in Deutschland etablierten Algorithmen, Richtlinien und Akteuren darstellt. Insbesondere bei den Softwarebeschreibungen wird deutlich, dass oftmals Assistenzsysteme zur Anwendung kommen. Darüber hinaus sind Studien von Interesse, welche den konkreten Einsatz von Softwaresystemen zur Entscheidungsfindung beschreiben. Basierend auf den vorliegenden Recherchen wird dies insbesondere im Personal- und Bewerbungsmanagement eingesetzt. Hunkenschroer und Luetge (2022) betrachten beispielsweise 51 Studien mit Bezug zum Personalmanagement. Sie konstatieren, dass eine normative Einschätzung der Algorithmen bislang im wissenschaftlichen Diskurs unterrepräsentiert ist. Darüber hinaus schlussfolgern sie, dass die Diskussion zu ethischen Herausforderungen zu stark an Richtlinien ausgerichtet ist und nicht konkret auf spezifische Anwendungsfelder bezogen werden.

Eine weitere Überblicksstudie in diesem Bereich von Will, Krpan und Lordan (2023) untersucht, inwiefern Menschen oder Algorithmen hinsichtlich der Entscheidungsfindung besser, gleich oder schlechter handeln. Sie analysieren dazu die Effizienz, Performanz, Diversität und wie die Entscheidungsfindung durch einen Algorithmus wahrgenommen wird. In ihrer Zusammenfassung zeigen die Autoren auf, dass insbesondere bezogen auf Effizienz und Performanz die Algorithmen einen Vorteil haben und auch die Diversität besser gesichert wird. Die Wahrnehmung der Algorithmen-Entscheidungen wird

jedoch generell durch Menschen als kritisch eingeschätzt, was gegebenenfalls an fehlenden Erklärungsansätzen zur Funktion der Software oder der Entscheidungsfindung liegt.

Der Einsatz von KI in ethischen Entscheidungssituationen ist also nicht einfach. Bei allen Problemen, die damit verbunden sind, legen die Studien allerdings auch nahe, dass sie in Form von Assistenzsystemen sehr wohl Entscheidungsträger*innen in konfliktiven Situationen unterstützen können. Sie können beispielsweise Muster in großen Datenmengen erkennen, die für Menschen schwer zu erfassen sind, und so wertvolle Informationen für die Entscheidungsfindung liefern. Dabei ist es wichtig, die Rolle von Algorithmen als Unterstützung und Ergänzung menschlicher Expertise zu verstehen, anstatt sie als Ersatz zu betrachten. Denn menschliche Fähigkeiten wie Empathie, Intuition und Urteilsvermögen bleiben bei ethischen Entscheidungen zentral, da diese von Algorithmen nur schwer nachgebildet werden können. Daher sollten Algorithmen darauf abzielen, die menschliche Entscheidungsfindung zu verbessern, anstatt sie zu ersetzen, und die Transparenz und Nachvollziehbarkeit algorithmischer Entscheidungen sicherstellen. Dies beinhaltet auch die Implementierung von Mechanismen zur Überprüfung und Anpassung von Algorithmen, um sicherzustellen, dass sie ethischen Standards entsprechen und kontinuierlich verbessert werden.

Eine solche Strategie ethischer Verantwortung im Kontext von KI-Entwicklungen ist sinnvoll, da das Hauptproblem auf Algorithmen basierten Entscheidungssystemen darin besteht, dass sie einen gültigen Wert benötigen, um ein Ergebnis zu erzeugen. Fehlende Werte müssen interpoliert werden, was die Ergebnisse verfälschen kann. Eine Vielzahl von Studien beschäftigt sich mit den daraus resultierenden Effekten, z.B. Fragen der Gleichbehandlung verschiedener Gruppen. Dabei zeigt sich jedoch, dass es unter Umständen keine wirklich fairen Algorithmen geben kann (Kleinberg et al. 2016) oder nur ein zeitlich nachgelagerter Auswertungsschritt helfen kann, fehlende Daten zu korrigieren (Chouldechova 2016). Zu diesem Zeitpunkt ist die Entscheidung des Algorithmus jedoch längst gefallen, woraus wieder (problematische) gesellschaftlichen Folgen resultieren können.

Um eine Antwort auf dieses Problem zu finden, wurde im Rahmen des KAI-Mo-Projektes ein Drei-Agenten-Lösungsansatz entwickelt, auf den im weiteren Verlauf noch detailliert eingegangen wird. Ziel ist es, dass das *Agentenarray* aus unabhängigen Ansätzen Lösungen entwickelt und der eingeschlagene Weg transparent dokumentiert wird.

Damit erfüllt ein solcher Ansatz wesentliche Forderungen, die in verschiedenen Fachgremien zur Thematik der ethisch-moralischen Herausforderungen von KI und deren Einsatz in einem digitalen gesellschaftlichen Kontext diskutiert werden. So definiert beispielsweise die *High-level expert group on artificial intelligence* der Europäischen Kommission in ihrer *Assessment List for Trustworthy AI* (ALTAI) (AI HLEG 2020) sieben Punkte: (1) menschliches Handeln und Kontrolle, (2) technische Robustheit und Sicherheit, (3) Datenschutz und Datenmanagement, (4) Transparenz, (5) Vielfalt, Nichtdiskriminierung und Fairness, (6) Umwelt- und gesellschaftliches Wohlergehen, (7) Verantwortlichkeit. Der Deutsche Ethikrat wiederum bezeichnet Entscheidungsunterstützungssysteme in der Stellungnahme *Mensch und Maschine – Herausforderungen durch Künstliche Intelligenz* (Deutscher Ethikrat 2023) als begrüßenswert, um die menschliche Entscheidungsfindung durch einen verbesserten Zugang zu Daten zu unterstützen. Eingeschränkt wird die Empfehlung jedoch durch den Hinweis, dass diese Systeme stets das Ziel verfolgen müssen, die menschliche Entscheidungsfindung zu verbessern und nicht die Effizienzsteigerung oder Personaleinsparung durch den Einsatz von Computersystemen voranzutreiben. Darüber hinaus wird in der Stellungnahme auf die potentielle Gefahr hingewiesen, dass Menschen eine Computerentscheidung unreflektiert übernehmen. Folgerichtig wird in diesem Zusammenhang empfohlen, dass »geeignete technische und organisatorische Instrumente zur Vorkehrung gegen die manifeste Gefahr eines Automation Bias bereitgestellt werden, die es den Fachkräften erschweren […] der algorithmischen Entscheidungsempfehlung unbesehen zu folgen.« (Deutscher Ethikrat 2023: 249) Diese Bedingung gilt für alle gesellschaftlichen Felder, in denen KI-Systeme zum Einsatz kommen. Es geht in diesem Zusammenhang beispielsweise um die Notwendigkeit, den Datenschutz und die Privatsphäre der betroffenen Personen zu wahren, sowie die Sicherstellung, dass die Algorithmen (beispielsweise durch die Zusammenstellung der Trainingsdaten) bestehende soziale Ungleichheiten nicht verstärken.

Dementsprechend ist es wichtig, die Ergebnisse und Anforderungen der Informatik in Bezug auf ethische Algorithmen kontinuierlich zu evaluieren und in den Dialog mit den anderen beteiligten Disziplinen einzubringen. Hierzu gehört auch die Bereitschaft, die entwickelten Systeme und Ansätze kritisch zu hinterfragen und gegebenenfalls anzupassen, um die bestmögliche Unterstützung für die Entscheidungsfindung in der Sozialen Arbeit zu gewährleisten.

Digitale Assistenzsysteme in der Sozialen Arbeit – ein Beispielfeld

In der Sozialen Arbeit ist die Entscheidungsunterstützung durch KI grundsätzlich ein sehr kontroverses Thema. Anhand der Debatte um die Entscheidungsfindung zur Kindeswohlgefährdung kann dies besonders deutlich illustriert werden, da hier die Fachkräfte im Jugendamt gewichtige Entscheidungen über die Zukunft junger Menschen und ihrer Familien treffen müssen (Bathke et al. 2019).

Schon lange vor der Entwicklung von Systemen der KI in diesem Bereich war die Soziale Arbeit auf der Suche nach Instrumenten, die die Prognosequalität in Kinderschutzfällen verbessern könnten. Dabei waren zwei große Linien zu erkennen: Einerseits zeigte sich in der Praxis, dass Fachkräfte mit Erfahrung deutlich bessere Prognosen generieren, als Personen, die neu in dem Feld arbeiten. Die große Bedeutung des Erfahrungswissens wird nicht zuletzt im §8a SGB VIII deutlich, der vorschreibt, dass Träger, die Leistungen im Kinder- und Jugendbereich erbringen, eine »insofern erfahrene Fachkraft« (§8a SGB VIII, Abs. 4, Nr. 2) bei Fällen einer vermuteten Kindeswohlgefährdung hinzuziehen müssen.

Andererseits spielen neben dem individuellen Erfahrungswissen in Fällen der Kindeswohlgefährdung wissenschaftlich abgesicherte Erkenntnisse eine besondere Rolle. In umfangreichen Studien konnten empirisch Faktoren herausgearbeitet werden, die das Risiko einer Kindeswohlgefährdung erhöhen oder vermindern. Aufbauend auf diesen Ergebnissen wurden mit statistischen Verfahren Modelle entwickelt, die unter Berücksichtigung der relevanten Faktoren mit transparenten Algorithmen Prognosen zur Risikostruktur liefern. Für die Praxis der Sozialen Arbeit wurden sie in Checklisten übersetzt, die teilweise in Papierform, teilweise digital bearbeitet werden können (Ackermann 2020).

Es ist sehr bemerkenswert, dass einerseits in vielen methodisch hochwertigen Metastudien herausgearbeitet wurde, dass diese Instrumente eine deutlich höhere Prognosequalität aufweisen als Fachkräfte, die intuitiv-diskursiv agieren (Grove et al. 2000; van der Put et al. 2017), andererseits im Fachdiskurs dennoch der Standpunkt vertreten wird, dass Ansätze, die eine individuelle Expertise der Fachkräfte in den Vordergrund stellen, zu favorisieren seien (Bastian 2012; Schroth 2021). Hier wird deutlich, dass in der Disziplin der Sozialen Arbeit auf der inhaltlichen Ebene eine große Sorge vor einer De-Professionalisierung durch die Nutzung standardisierter Verfahren besteht (Schrödter et al. 2020). Insofern überrascht es nicht, dass die Anwendung von KI-Sys-

temen im deutschen Kinderschutz sehr kritisch diskutiert wird und die meisten Autor*innen eher die Risiken der Technologie betonen (Görder 2021).

Im internationalen Kontext liegen dagegen bereits einige Ansätze vor, die größere Datenbestände und komplexere Algorithmen für die Risikoprognostik nutzen (La Valle et al. 2016). In Neuseeland wurde z.B. ein *Predictive Risk Modelling to Prevent Child Maltreatment* (PRM) entwickelt. Hier wurden Daten aus den verschiedenen staatlichen Systemen, wie der Sozial- und Kinderfürsorge und dem Gesundheits- und Erziehungssystem verknüpft und darauf aufbauend ein Risikoscore ermittelt (Gillingham 2021). Einen ähnlichen Ansatz verfolgen die Kinderschutzbehörden in den USA. Dort wird z.B. im *Allegheny Family Screening Tool* (AFST) eine Verknüpfung unterschiedlicher Datenquellen zur Berechnung eines *Familien-Screening-Scores* verwendet. Bei einem hohen Score-Wert wird ein höheres Risiko der Kindeswohlgefährdung angenommen und entsprechende Interventionen eingeleitet (Holstein 2022).

Die praktische Anwendung dieser Verfahren zeigt jedoch einige grundlegende Probleme, die mit den skizzierten ethischen Bedenken in Zusammenhang stehen. So zeigte sich, dass das PRM einen sehr starken Zusammenhang zwischen Armut und dem Merkmal ›Alleinerziehend‹ und dem Risiko einer Kindeswohlgefährdung herstellt. Dies lässt den Rückschluss zu, dass in den verwendeten Daten Fälle mit dieser Kombination von Merkmalen häufig vorkamen. Die Interpretation, dass Armut und Alleinerziehend ursächlich für die Kindeswohlgefährdung sind, ist allerdings eine Verwechslung von Korrelation und Kausalität. Aufgrund dieser Problematik wurden Entwicklung und Implementation dieses Verfahrens eingestellt. Ähnliche Vorwürfe wurden gegen das AFST vorgebracht, das ebenfalls vor allem als Armuts-Profiling funktioniert hat (Eubanks 2018).

Die internationalen Erfahrungen machen deutlich, dass die Einführung KI-basierter Risikoprognoseverfahren mit vielfältigen Problemen einhergehen. Eine besondere Rolle kommt den verwendeten Trainingsdaten zu. In den genannten Anwendungen wurden die Algorithmen mit Datensätzen trainiert, in denen durch die Auswahl der betroffenen Personen oder durch implizite Stereotype bei den bearbeitenden Personen Verzerrungen von den Algorithmen erlernt wurden. Bei zukünftigen Entwicklungen ist es daher entscheidend, schon bei der Wahl des Trainingsmaterials auf die Einhaltung von Gerechtigkeitsprinzipien und eine diskriminierungsfreie Auswahl der Daten zu achten (Görder 2021).

Für die Soziale Arbeit in Deutschland stellt sich die Frage, wie dieses Trainingsmaterial aufgebaut sein müsste, um Algorithmen zu trainieren, die Er-

gebnisse produzieren, die sowohl frei von Diskriminierung als auch mit hoher Vorhersagequalität ausgestattet sind. Ausgehend von dem bestehenden Widerspruch im Fachdiskurs zwischen statistischen Prognosemodellen und erfahrungsbasierten Ansätzen könnte der Einsatz von KI-Systemen hier zu einer Art Synthese führen. Bisher bestand bei der Nutzbarmachung des Erfahrungswissens der Fachkräfte das Problem, dass das zugrunde liegende implizite Wissen nur schwer explizierbar war (Böhle 2020). Da dieses Wissen jedoch die Grundlage von Entscheidungen darstellt, die in Akten dokumentiert sind, wird die Nutzung solcher prozessgenerierten Textdokumente in der Professionsforschung der Sozialen Arbeit schon länger diskutiert.

Mit den klassischen qualitativen Analyseverfahren konnten hier bisher keine umfangreichen Studien durchgeführt werden (Lehmann und Klug 2019). Verfahren aus dem KI-Forschungsbereich können neue Erkenntnisse generieren. So könnten die großen Textmengen, die in deutschen Jugendämtern vorliegen mit den verschiedenen Verfahren untersucht und entsprechende Muster aus den Daten extrahiert werden. Es ist anzunehmen, dass das bisher implizite Erfahrungswissen der Fachkräfte auf diese Weise zumindest ansatzweise verarbeitet und zum Ausdruck gebracht werden könnte (Vladova et al. 2019). Es besteht eine große Chance, mit maschinellen Lernverfahren aus der dokumentierten Fachlichkeit in Akten die Grundlagen der Entscheidungsfindung erfahrener Fachkräfte zu erlernen und darauf aufbauend Assistenzsysteme zu generieren.

Auch wenn dieser Ansatz zunächst sehr naheliegend erscheint, birgt er jedoch ebenfalls große Risiken. So sind in den Akten der deutschen Jugendämter ähnlich wie in USA und Neuseeland bestimmte Bevölkerungsgruppen überrepräsentiert (Jugendinstitut eV et al. 2020). Auch die Arbeit im Jugendamt vor Ort ist nicht frei von Vorurteilen und Ressentiments, die sich in der fachlichen Entscheidungsfindung widerspiegeln (Harrer-Amersdorffer 2022; Herzog 2022). Es kann also nicht empfohlen werden, bestehende Daten unreflektiert zum Training einer KI zur Entscheidungsunterstützung zu nutzen. Als Vorstufe wäre eine Mustererkennung durch maschinelle Lernverfahren in den Dokumenten bestehender Praxis allerdings sehr hilfreich. Die erkannten Muster würden sichtbar und könnten damit Gegenstand einer fachlichen Debatte sein. Aufbauend auf einer in diesem Sinne von sozialarbeiterischer Fachlichkeit geprägten Entwicklungsarbeit ist ein System zur Entscheidungsunterstützung in Kinderschutzfällen sehr wünschenswert und denkbar (Linnemann et al. 2023; Steiner und Tschopp 2022).

Abgesehen von der inhaltlichen Qualität von KI-gestützten Assistenzsystemen stellt sich die Frage ihrer Einbindung in die Entscheidungspraktiken der Jugendämter. Selbst wenn ein solches Verfahren extrem gute Prognoseergebnisse liefern sollte, sehen einerseits die geltende Rechtslage, andererseits auch tiefgreifende ethische Erwägungen (Deutscher Ethikrat 2023) vor, dass so lebensentscheidende Entscheidungen in Letztverantwortung von Menschen getroffen werden müssen. Daher muss auch hier überlegt werden, wie eine sinnvolle Integration in Entscheidungsprozesse aussehen kann. Dabei ist im Kontext der Sozialen Arbeit zu beachten, dass hier seit langem eine hohe Skepsis gegenüber digitaler Technologie vorliegt (Bertsche und Como-Zipfel 2017), sodass ein entsprechendes Unterstützungssystem evtl. weniger stark genutzt werden würde als in anderen Bereichen. Gleichzeitig ist bekannt, dass Menschen dazu neigen, ihrer eigenen Expertise weniger zu vertrauen, als einem maschinellen System und im Zweifel eher der Einschätzung einer Maschine folgen, selbst wenn sie dem Urteil der Maschine nicht völlig zustimmen (Banbury 2021; Gapski 2020). Daher besteht das Risiko, dass einem KI-System zu viel Vertrauen entgegengebracht wird. Die Einbindung eines KI-Systems in einen Entscheidungsfindungsprozess muss also sowohl sicherstellen, dass die Expertise des Systems in ausreichender Intensität berücksichtigt wird, als auch verhindern, dass die Menschen unreflektiert der Einschätzung des technischen Systems folgen.

Aus der derzeitigen Perspektive sind daher einfache Empfehlungssysteme im Kinderschutz, die sich deutlich für bestimmte Maßnahmen aussprechen, sehr kritisch zu sehen. Aktuell existiert im deutschsprachigen Raum kein KI-System, das mit einem Datensatz trainiert wurde, der den Ansprüchen an die inhaltliche Qualität und Diskriminierungsfreiheit genügt. Weiterhin liegt in der Interpretation der Ergebnisse und der Interaktion zwischen Mensch und Maschine aktuell ein großes Fehlerpotenzial. Daher erscheint es sinnvoll, einerseits die Entwicklung fachlich kontrollierter und diskriminierungsfreier Systeme voranzutreiben und andererseits Konzepte zu entwickeln, die eine zielführende Interaktion von Mensch und Maschine ermöglichen. Ein Ansatz könnte eine sehr defensive Integration einer KI sein, die nie eine eigene Empfehlung ausspricht, sondern den Menschen bei seiner Entscheidungsfindung durch gut aufbereitete Informationen und fachliche Hinweise maßgeblich unterstützt.

Ausblick

Vor dem Hintergrund der skizzierten Problemstellung versammelt der Band Beiträge aus dem Feld der Sozialen Arbeit – aber auch anderen Themenfeldern, wie u.a. der Medizin oder der Politik. Die Beiträge rekonstruieren soziale Praktiken und Institutionen, in denen Algorithmen bereits zum Einsatz kommen oder dies für die Zukunft geplant ist. In dem Band bieten Expert*innen aus verschiedenen Fachrichtungen fundierte Einsichten in die KI-gestützte Entscheidungs- und Urteilsfindung. Von der digitalen Operationalisierung über die Rolle des Menschen im Zentrum des technischen Fortschritts bis hin zur Konzeption von vertrauenswürdigen Systemen. Anhand dieser Beispiele werden sowohl Chancen als auch Grenzen des Einsatzes von KI-Systemen diskutiert, v.a. in hoch konfliktiven Situationen. Es geht letztlich um die Frage nach der ethischen Verantwortung im digitalen Zeitalter. Für die Diskussion dieser Frage will der Band einige grundlegende Impulse geben.

Literatur

Ackermann, Timo. 2020. »Digitalisierung in der Kinder- und Jugendhilfe und im Kinderschutz: Von Risikoeinschätzungsbögen über Fallbearbeitungssoftware bis zu Big Data.« *Soziale Passagen*, 12 (1): 171–177.

AI HLEG. 2020. »Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment | Shaping Europe's digital future.« Accessed October 6, 2023. https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment.

Awad, Edmond, Sohan Dsouza, Richard Kim, Jonathan Schulz, Joseph Henrich, Azim Shariff et al. 2018. »The Moral Machine experiment.« *Nature* 563 (7729): 59–64. https://doi.org/10.1038/s41586-018-0637-6.

Bastian, Pascal. 2012. »Die Überlegenheit statistischer Urteilsbildung im Kinderschutz–Plädoyer für einen Perspektivwechsel hin zu einer angemessenen Form sozialpädagogischer Diagnosen.« In *Rationalitäten des Kinderschutzes,* edited by Thomas Marthaler, Pascal Bastian, Ingo Bode, Mark Schrödter, 249–267. Wiesbaden: Springer.

Bathke, Sigrid A., Milena Bücken, Dirk Fiegenbaum. 2019. »Die Grundlagen: Kinderschutz, Kindeswohl und Kindeswohlgefährdung aus rechtlicher und fachlicher Perspektive.« In *Praxisbuch Kinderschutz interdisziplinär,*

edited by Sigrid A. Bathke, Milena Bücken, Dirk Fiegenbaum, 5–106. Wiesbaden: Springer VS.

Bertsche, Oliver, Frank Como-Zipfel. 2017. »Sozialpädagogische Perspektiven auf die Digitalisierung.« *Soziale Passagen*, 8(2): 235–254. https://doi.org/10.1007/s12592-016-0244-z.

Beuth, Patrick, Heiner Hoffmann, Max Hoppenstedt. 2023. »Das sind die Menschen hinter der KI-Revolution.« Accessed October 6, 2023. https://www.spiegel.de/netzwelt/web/clickwork-und-content-moderation-die-gesichter-hinter-der-kuenstlichen-intelligenz-a-9629ea15-5bc3-42bd-a236-199d606b1a24.

Böhle, Fritz. 2020. »Implizites Wissen und subjektivierendes Handeln – Konzepte und empirische Befunde aus der Arbeitsforschung.« In *Implizites Wissen: Berufs- und wirtschaftspädagogische Annäherungen*, edited by Rico Hermkes, Georg Hans Neuweg, Tim Bonowski, 36–64. Bielefeld: Wbv.

Brieger, Julchen. 2018. »Über die Unmöglichkeit einer kantisch handelnden Maschine.« In *Maschinenethik. Normative Grenzen autonomer Systeme,* edited by Matthias Rath, Friedrich Krotz und Matthias Karmasin, 107–120. Wiesbaden: Springer Fachmedien.

Chouldechova, Alexandra. 2016. »Fair prediction with disparate impact: A study of bias in recidivism prediction instruments.« https://doi.org/10.48550/ARXIV.1610.07524.

Christen, Markus et al. 2020. »Wenn Algorithmen für uns entscheiden: Chancen und Risiken der künstlichen Intelligenz.« Zürich: Vdf Hochschulverlag AG an der ETH Zürich.

Deutscher Ethikrat. 2023. »Mensch und Maschine – Herausforderungen durch Künstliche Intelligenz.« Accessed October 6, 2023. https://www.ethikrat.org/fileadmin/Publikationen/Stellungnahmen/deutsch/stellungnahme-mensch-und-maschine.pdf.

Dietrich, Eric. 2011. »Homo Sapiens 2.0. Building the Better Robots of Our Nature.« In *Machine ethics,* edited by Michael Anderson, Susan Leigh Anderson, 531–538. Cambridge: Cambridge university press.

Dignum, Virginia. 2018. »Ethics in artificial intelligence: introduction to the special issue.« *Ethics and information technology* 20 (1): 1–3. https://doi.org/10.1007/s10676-018-9450-z.

Eubanks, Virginia. 2018. *Automating inequality: How high-tech tools profile, police, and punish the poor*. New York: St. Martin's Press.

Floridi, Luciano, John W. Sanders. 2004. »On the Morality of Artificial Agents.« *Minds and Machines* 14 (3): 349–379. https://doi.org/10.1023/B:MIND.0000035461.63578.9d.

Gapski, Harald. 2020. »Digitale Transformation: Datafizierung und Algorithmisierung von Lebens-und Arbeitswelten.« In *Handbuch Soziale Arbeit und Digitalisierung*, edited by Nadia Kutscher, Thomas Ley, Udo Seelmeyer, Friederike Siller, Angela Tillmann, Isabel Zorn, 156–166. Weinheim: Beltz Juventa.

Gillingham, Philip. 2021. »Practitioner perspectives on the implementation of an electronic information system to enforce practice standards in England.« *European Journal of Social Work 24* (5): 761–771. https://doi.org/10.1080/13691457.2020.1870213.

Görder, Björn. 2021. »Die Macht der Muster. Die Ethik der Sozialen Arbeit vor professionsbezogenen und gesellschaftlichen Herausforderungen durch, künstliche Intelligenz ›.« *Ethik Journal* 7 (2). Accessed October 6, 2023. https://www.ethikjournal.de/fileadmin/user_upload/ethikjournal/Texte_Ausgabe_2021_2/Goerder_Ethikjournal_2.2021.pdf.

Grove, William M., David H. Zald, Boyd S. Lebow, Beth E. Snitz, Chad Nelson. 2000. »Clinical versus mechanical prediction: A meta-analysis.« *Psychological Assessment*, *12* (1): 19–30. https://doi.org/10.1037/1040-3590.12.1.19.

Gutwald, Rebecca, Michael Reder. 2023. »How to Protect Children? A Pragmatic Approach: On State Intervention and Children's Welfare.« *The Journal of Ethics* 27 (1): 77–95. https://doi.org/10.1007/s10892-022-09416-3.

Gutwald, Rebecca, Jennifer Burghardt, Maximilian, Kraus, Michael Reder, Robert Lehmann, Nicholas Müller. 2021. »Soziale Konflikte und Digitalisierung. Chancen und Risiken digitaler Technologien bei der Einschätzung von Kindeswohlgefährdungen.« *Ethik Journal* 7 (2). Accessed October 6, 2023. https://www.ethikjournal.de/fileadmin/user_upload/ethikjournal/Texte_Ausgabe_2021_2/Gutwald_u.a._Ethikjournal_2.2021.pdf.

Harrer-Amersdorffer, Jutta. 2022. *Fachliches Handeln in der Fallarbeit: Eine empirische Studie über den Stand der Sozialpädagogischen Familienhilfe*. Leverkusen: Verlag Barbara Budrich.

Herzog, Lucas-Johannes. 2022. »Rassismus im Jugendamt: Vom Nachdenken über eine nicht geführte Debatte.« *Forum Erziehungshilfen*, 10 (1): 11–13. https://doi.org/10.3262/FOE2201011.

Holstein, Kenneth. 2022. »What happens when human workers oversee algorithmic tools?« Accessed October 6, 2023. https://medium.com/@ken

neth.holstein/what-happens-when-human-workers-oversee-algorithmic-tools-bbfc32e8ce61.

Hunkenschroer, Anna Lena, Christoph Luetge. 2022. »Ethics of AI-Enabled Recruiting and Selection: A Review and Research Agenda.« *Journal of Business Ethics*, 178 (4): 977–1007. https://doi.org/10.1007/s10551-022-05049-6.

Jaume-Palasí, Lorena, Matthias Spielkamp. 2017. »Ethik und algorithmische Prozesse zur Entscheidungsfindung oder -vorbereitung.« Accessed October 6, 2023. https://algorithmwatch.org/de/wp-content/uploads/2017/06/AlgorithmWatch_Arbeitspapier_4_Ethik_und_Algorithmen.pdf.

Jugendinstitut e.V., Deutsche, Susanne Lochner, Alexandra Jähnert. 2020. *DJI-Kinder-und Jugendmigrationsreport 2020: Datenanalyse zur Situation junger Menschen in Deutschland*. Bielefeld: Wbv.

Kleinberg, Jon, Sendi Mullainathan, Manish Raghavan. 2016. »Inherent Trade-Offs in the Fair Determination of Risk Scores.« https://doi.org/10.48550/ARXIV.1609.05807.

Kucklick, Christoph. 2016. »Soziologische Aspekte von Big Data.« Audioprotokoll, Deutscher Ethikrat, March 23, 2016. https://www.ethikrat.org/sitzungen/2016/big-data.

La Valle, Ivana, Berni Graham, Lisa Payne. 2016. »*A consistent identifier in education and children's services*.« Department for Education. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/534744/Consistent_identifier__report_July_2016.pdf.

Lehmann, Robert, Wolfgang Klug. 2019. »Die prozessorientierte Aktenanalyse.« In *Sekundäranalysen in der Kinder- und Jugendhilfe* edited by Maik-Carsten Begemann, Klaus Birkelbach, 301–319. Wiesbaden: Springer.

Linnemann, Gesa Alena, Julian Löhe, Beate Rottkemper. 2023. »Bedeutung von Künstlicher Intelligenz in der Sozialen Arbeit.« *Soziale Passagen* 15: 197–211. https://doi.org/10.1007/s12592-023-00455-7.

Lukesch, Helmut. 2006. FEPAA. Fragebogen zur Erfassung von Empathie, Prosozialität, Aggressionsbereitschaft und aggressivem Verhalten. Göttingen: Hogrefe.

Matzat, Lorenz, Lukas Zielinski, Miriam Cocco, Kristina Penner, Matthias Spielkamp, Sebastian Gießler, Sebastian Lang, Veronika Thiel. 2019. »Atlas der Automatisierung/Automatisierte Entscheidungen und Teilhabe in Deutschland.« https://atlas.algorithmwatch.org/wp-content/uploads/2019/07/Atlas_der_Automatisierung_von_AlgorithmWatch.pdf.

Misselhorn, Catrin. 2018a. Grundfragen der Maschinenethik. Ditzingen: Reclam.

Misselhorn, Catrin. 2018b. »Können und sollen Maschinen moralisch handeln?« *APuZ* 68 (6–8): 29–33.

Rutherford, Nichola. 2023. »Drake and The Weeknd AI song pulled from Spotify and Apple.« BBC News. Accessed November 9, 2023. https://www.bbc.com/news/entertainment-arts-65309313.

Schrödter, Mark, Pascal Bastian, Brian Taylor. 2020. »Risikodiagnostik und Big Data Analytics in der Sozialen Arbeit.« In *Handbuch Soziale Arbeit und Digitalisierung*, edited by Nadia Kutscher, Thomas Ley, Udo Seelmeyer, Friederike Siller, Angela Tillmann, Isabel Zorn, 255–264. Weinheim: Beltz-Juventa.

Schroth, Emma. 2021. »Digitale Falldokumentation im Jugendamt.« *Sozial Extra*, *45*(1): 49–52. https://doi.org/10.1007/s12054-020-00350-y.

Shevat, Amir. 2017. Designing Bots—Creating Conversational Experiences. Sebastopol: O'Reilly.

Steiner, Oliver, Dominik Tschopp. 2022. »Künstliche Intelligenz in der Sozialen Arbeit.« *Sozial Extra*, *46*(6): 466–471. https://doi.org/10.1007/s12054-022-00546-4.

van der Put, Claudia E., Mark Assink, Noëlle F Boekhout van Solinge. 2017. »Predicting child maltreatment: A meta-analysis of the predictive validity of risk assessment instruments.« *Child abuse & neglect*, *73*: 71–88. https://doi.org/10.1016/j.chiabu.2017.09.016.

van Wynsberghe, Aimee, Scott Robbins. 2019. »Critiquing the Reasons for Making Artificial Moral Agents.« *Science and engineering ethics* 25 (3): 719–735. https://doi.org/10.1007/s11948-018-0030-8.

Vladova, Gergana, Norbert Gronau, Leo Sylvio Rüdian. 2019. »Wissenstransfer in Bildung und Weiterbildung: Der Beitrag Künstlicher Intelligenz.« In *Digitale Transformation – Gutes Arbeiten und Qualifizierung Aktiv Gestalten*, edited by Dieter Spath, Birgit Spanner-Ulmer, 89–106. Berlin: Gito-Verlag.

Waldrop, Mitchell. 1987. »A question of responsibility.« *The AI Magazine* 8 (1): 28. https://doi.org/10.1609/aimag.v8i1.572.

Wallach, Wendell, Colin Allen. 2009. Moral machines. Teaching robots right from wrong. Oxford: Oxford University Press.

Weber, Karsten. 2018. »Autonomie und Moralität als Zuschreibung. Über die begriffliche und inhaltliche Sinnlosigkeit einer Maschinenethik.« *Maschinenethik. Normative Grenzen autonomer Systeme*, edited by Matthias Rath, Friedrich Krotz und Matthias Karmasin, 193–210. Wiesbaden: Springer Fachmedien.

Wickens, Christopher D., William S. Helton, Justin G. Hollands, Simon Banbury. 2021. *Engineering Psychology and Human Performance*. New York: Routledge.

Will, Paris, Dario Krpan, Grace Lordan. 2023. »People versus machines: Introducing the HIRE framework.« *Artificial Intelligence Review* 56 (2): 1071–1100. https://doi.org/10.1007/s10462-022-10193-6.

Xie, Jiangjian, Yujie Zhong, Junguo Zhang, Shuo Liu, Changqing Ding, Andreas Triantafyllopoulos. 2023. »A review of automatic recognition technology for bird vocalizations in the deep learning era.« *Ecological Informatics* no. 73(March): 101927. https://doi.org/10.1016/j.ecoinf.2022.101927.

Zweig, Katharina Anna, Georg Wenzelburger, Tobias D. Krafft. 2018. »On Chances and Risks of Security Related Algorithmic Decision Making Systems.« *European Journal for Security Research* 3(1): 181–203. https://doi.org/10.1007/s41125-018-0031-2.

Geist, Intelligenz, Information und Daten – Artificial Intelligence im Wandel der Wissenschaftskulturen

Eine ideengeschichtliche Begriffsverortung

Rudolf Seising

Zu welcher Wissenschaftsdisziplin oder -kultur sollten die Forschungen zur Artificial Intelligence (AI) gezählt werden? Wie passen sie zu den Geistes-, Natur- oder Ingenieurwissenschaften, den Kultur-, Sozial- und Strukturwissenschaften? Die historische Betrachtung findet zahlreiche Ansätze zur AI-Forschung in verschiedenen wissenschaftlichen Disziplinen, allen voran die in den USA in den 1950er Jahren entstandenen Computer Sciences bzw. einige Jahre später in Europa die Informatik. Auch diese Fächer waren nicht aus einem Guss entstanden, wie 1971 der österreichische Computer-Pionier Heinz Zemanek (1920–2014) schrieb:

> »Mathematik und Nachrichtentechnik, Buchhaltung und Statistik sind zwar Wurzeln und Bausteine, aber seit geraumer Zeit bilden sie nicht mehr den Kern der Computer-Wissenschaften, und nichts wäre verkehrter, als die Informatik als Konglomerat der eben genannten Felder zu konzipieren; was von ihnen noch bleiben wird in der Informatik, muss sehr kritisch geprüft sein.« (Zemanek 1971, 158)

Weiter nannte Zemanek die Informatiker hier »Ingenieure abstrakter Objekte« und 1974 definierte der Münchner TU-Professor Friedrich L. Bauer (1924–2015) die Informatik als »Ingenieur-Geistes- bzw. Geistes-Ingenieurwissenschaft«.[1] Sie sei die »Wissenschaft von der Programmierung der

1 Dieser Ausspruch inspirierte zum Namen des von 2019 bis 2023 vom BMBF geförderten wissenschafts- und technikhistorischen Forschungsprojekts »IGGI – Ingenieur-Geist und Geistes-Ingenieure: Eine Geschichte der Künstlichen Intelligenz in der

Informations- das heißt Zeichenverarbeitung« und ihre Resultate seien die geschriebenen Programme, die nicht materiell, sondern immaterielle, abstrakte Objekte sind.

Jüngere Wissenschaftler*innen in der Bundesrepublik Deutschland, die an den Forschungen der AI interessiert waren, wollten ihre Fächer für solche Untersuchungen öffnen. So gab es Keimzellen für die AI-Forschung nicht nur in Elektrotechnik und Mathematik, sondern auch in der Linguistik, in der Philosophie, in Psychologie und Soziologie und in den Wirtschaftswissenschaften. In einem 1966 von Kommunikations- bzw. Sprachwissenschaftlern der Universität Bonn verfassten Überblicksbericht zum damaligen Stand der AI-Forschung war »Künstliche Intelligenz« noch in Anführungszeichen geschrieben, denn dies sei

> »kein wohldefinierter, einheitlich verwendeter Terminus. Zuerst geprägt im Jargon der Fachwissenschaftler, die sich mit dem Einsatz von Computern bei nicht-numerischen Problemen beschäftigten, hat er eher den Charakter eines Schlagwortes.« (Ungeheuer, Krallmann, Schnelle, and Tillmann 1966, 1)

Je nach Ort, Zeit, Kultur und Gesellschaft lassen sich verschiedene Geschichten der AI-Forschung erzählen.[2] Ein Kenner dieser Narrative ist der Informatiker Wolfgang Bibel (* 1938),[3] der gemeinsam mit Ulrich Furbach (* 1948) rückblickend neben der weitgehenden Unkenntnis »der Grundlagen für den universellen Berechenbarkeitsbegriff« auf deutscher Seite im Gegensatz zur angelsächsischen einen weiteren Grund »in der jeweiligen Rolle der Geisteswissenschaften« sieht:

> »Bekanntlich übt Wilhelm Dilthey bis heute auf diese akademischen Disziplinen, als deren Begründer er gilt, einen erheblichen Einfluss aus und dies

---

Bundesrepublik Deutschland« (Förderkennzeichen 01IS19029): https://www.deutsches-museum.de/forschung/forschungsinstitut/projekte/detailseite/iggi-ingenieur-geist-und-geistes-ingenieure.

2 Einige neuere AI-Geschichtsdarstellungen sind in dem Special Issue »Dynamics of AI: European Histories« der IEEE Annals of the History of Computing im letzten Heft des Jahres 2023 erschienen.

3 Bibel war von 1987 bis 1988 Professor an der University of British Columbia in Vancouver (Kanada) und danach Professor für Intellektik an der Technischen Universität Darmstadt (1988–2004). Er war im Vorstand der »International Joint Conferences on Artificial Intelligence« (IJCAI, 1986–1992) und von 1987 bis 1989 deren Präsident.

> besonders, aber nicht nur, im deutschsprachigen Raum, während sein Einfluss auf die angelsächsischen »humanities« wesentlich schwächer ausgeprägt ist.« (Bibel and Furbach 2018, 43)

Der Theologe und Philosoph Wilhelm Dilthey (1833–1911) hatte im Jahre 1883 eine »Einleitung in die Geisteswissenschaften« publiziert, die ihre Fortsetzung in seiner erstmals 1910 erschienenen Abhandlung »Der Aufbau der geschichtlichen Welt in den Geisteswissenschaften« fand, und deren spätere Auflagen jeweils noch erweitert erschienen, ohne dass das Werk zum Abschluss kam. Er führte den Begriff der Geisteswissenschaften ein als »[d]as Ganze der Wissenschaften, welche die geschichtlich-gesellschaftliche Wirklichkeit zu ihrem Gegenstande haben« (Dilthey 1883, 5). Es war der Versuch, für die geisteswissenschaftliche Erkenntnis eine logischen Grundlegung zu finden:

> »Neben den Naturwissenschaften hat sich eine Gruppe von Erkenntnissen entwickelt, naturwüchsig, aus den Aufgaben des Lebens selbst, welche durch die Gemeinsamkeit des Gegenstandes miteinander verbunden sind. Solche Wissenschaften sind Geschichte, Nationalökonomie, Rechts- und Staatswissenschaften, Religionswissenschaft, das Studium von Literatur und Dichtung, von Raumkunst und Musik, von philosophischen Weltanschauungen und Systemen, endlich die Psychologie. Alle diese Wissenschaften beziehen sich auf dieselbe große Tatsache: das Menschengeschlecht. Sie beschreiben und erzählen, urteilen und bilden Begriffe und Theorien in Beziehung auf diese Tatsache.« (Dilthey 1883, 3)

»Geist« waren für Dilthey alle Inhalte der menschlichen Kultur, die zwar das einzelne menschliche Bewusstsein übersteigen, wobei aber »das Individuum als Repräsentant der Gemeinsamkeit [dieser Kulturinhalte] erscheint«. (Dilthey 1883, 83) Den Geisteswissenschaften sei die Methode des Verstehens eigen, die der Theologe und Philosoph Friedrich Daniel Ernst Schleiermacher (1768–1834) für die Erschließung des Sinns von Texten begründete, die nicht nur Textauslegung bzw. -interpretation war, sondern deren Verstehen bedeutete. Schleiermacher nahm dafür die Bezeichnung »Hermeneutik« in Anspruch und ihm ging es dabei auch darum, die Probleme zu erfassen, die sich aus den Texten für Theologie, Rechts-, Literatur- und Geschichtswissenschaften ergaben. So wandelte sich die ursprünglich im Mittelalter lediglich als Textauslegung verstandene Hermeneutik in der Neuzeit zu einer allgemeinen Methodenlehre der Interpretation und des Verstehens. Dilthey legte

sie schließlich der Geisteswissenschaft zugrunde, die für ihn Erfahrungswissenschaft der »geistigen Erscheinungen« oder empirische Wissenschaft der »geistigen Welt« war (Dilthey 1910).

Eine Gegenüberstellung der Wissenschaften des Materiellen und des Immateriellen findet sich schon in der »Phänomenologie des Geistes« von Georg Wilhelm Friedrich Hegel (1770–1831) und danach in dessen »Wissenschaft der Logik«, wo er der Naturwissenschaft eine Wissenschaft des Geistes gegenüberstellte, die allerdings den »Geist« als eine ihrer selbst gewisse Vernunft zugrunde legt, aus dem heraus sich diese Wissenschaft deduktiv entwickelt. Hegels Begriff des Geistes war auf das Leben des Geistes in einer Gruppe, einem Volk oder einer Kultur geprägt. Seine Wissenschaft des Geistes betraf die vom Menschen selbst erzeugten »geisthaften Gegenständlichkeit« (Blasche 2008). Sie war keine Geisteswissenschaft im heutigen Sinne, denn sie suchte nicht menschlich-geschichtliche Verhältnisse mit den Mitteln der Empirie zu erfassen. Eine *empirische* Wissenschaft des Geistes, die auch historische und soziale Verhältnisse berücksichtigt, bedeutete die »Geisteswissenschaft« wohl erstmals, als der Chemiker Jakob Heinrich Wilhelm Schiel (1813–1889) das Wort verwendete, und zwar als Übersetzer der Arbeiten von John Stuart Mill (1806–1873). Dessen Werk »A System of Logic« behandelt im 6. Buch mit der Überschrift »Logic of Moral Sciences« logische und methodologische Probleme aus Psychologie, Soziologie, Ökonomie und Geschichte; in der deutschen Ausgabe heißt dieses Buch »Von der Logik der Geisteswissenschaften oder moralischen Wissenschaften«. Mill sah diese »moral sciences« in ihrer Entwicklung gegenüber den Naturwissenschaften zurückstehen und schlug vor, dass sie sich diese zum Vorbild nehmen sollten.

Wissenschaftskulturen

Als Mitglied des Wiener Kreises hatte Rudolf Carnap (1891–1970) einen logischen Aufbau der Welt beschrieben (Carnap 1928), Diltheys Aufbau galt der geschichtlichen Welt in den Geisteswissenschaften und nahm Bezug auf Hegels Aufbau der »geistigen Welt«; das war eine zweite Welt oder »zweite Natur« wie Theodor W. Adorno (1903–1969) formulierte (Adorno 1966, 289).

Nicht zwei Naturen sondern zwei Kulturen der Wissenschaft über die Natur postulierte der englische Physiker und Schriftsteller Charles Percy Snow (1905–1980), als er am 7. Mai 1959 im Senate House in Cambridge die Sir Robert Rede's Lecture unter dem Titel »The Two Cultures and the Scientific

Revolution« hielt. Die Kulturen der Naturwissenschaft und Technik (Sciences) einerseits und der Geisteswissenschaft und Literatur (Humanities) andererseits trenne nämlich eine Kluft: Naturwissenschaftler*innen und Ingenieur*innen einerseits und Geisteswissenschaftler*innen und Literat*innen andererseits analysieren, (re)konstruieren, erklären bzw. verstehen die Welt auf unterschiedliche Weisen. Die Kluft zwischen beiden »Denkwelten« sei fast unüberbrückbar geworden, es gebe kaum Kommunikation zwischen ihnen (Snow 1959). Snow erwartete eine Verarmung auf beiden Seiten, weil die damaligen Curricula keinerlei Platz für Interdisziplinarität ließen:

> »The clashing point of two subjects, two disciplines, two cultures -- of two galaxies, so far as that goes-ought to produce creative chances. In the history of mental activity that has been where some of the breakthroughs came. The chances are there now. But they are there, as it were, in a vacuum, because those in the two cultures can't talk to each other. It is bizarre how very little of twentieth-century science has been assimilated into twentieth-century art.« (Snow 1998, 16)

Vier Jahre später sah er das Entstehen einer »dritten Kultur« voraus, denn eine neue Generation von Naturwissenschaftler*innen werde die Kluft zwischen den beiden Kulturen überbrücken (Snow 1964). Seither wurde oft versucht, Bewegungen auszumachen, die die Unterschiede zwischen den beiden Kulturen ausgleichen. So veröffentlichte John Brockman 1995 den Sammelband »The Third Culture: Beyond the Scientific Revolution«, in dem er diejenigen Wissenschaftler*innen, die sich der Popularisierung von Wissenschaft in ihren Veröffentlichungen widmen und Antworten »auf die letzten Fragen« für ein breites Publikum geben, als eine solche »dritte Kultur« (Brockman 1995). Im Septemberheft 2007 des Scientific American diskutierte der Wissenschaftshistoriker und -journalist Michael Shermer sogenannte »integrative science«: wissenschaftliche Erzählungen »between technical and popular science writing«, »that blends data, theory and narrative«:

> »We are storytellers. If you cannot tell a good story about your data and theory – that is, if you cannot explain your observations, what view they are for or against and what service your efforts provide – then your science is incomplete.« (Shermer 2007)

»Dritte Kulturen« entstanden oftmals im Gefolge eines neuen Zugangs, einer neuen Theorie oder einer neuen Technologie. Im 20. Jahrhundert wurden natur- und geisteswissenschaftliche Argumentationen verknüpft, etwa wenn die quantenmechanische Unbestimmtheit als mögliche Erklärung für die Willensfreiheit des Menschen herangezogen wurde (Jordan 1932), oder wenn der genetische Reproduktionsprozess als Informationsübertragung gedeutet wurde (Kay 2002).

In seiner Besprechung von Snows »The Two Culture. And a Second Look« berichtete Ormsbee W. Robinson (1911–1995), der damals IBM's Direktor für die Planung der Hochschulbeziehungen war, dass das Symposium »The Impact of Science« im Jahre 1965 nur von 200 Teilnehmer*innen besucht wurde, während noch im Jahr zuvor, als das Symposium Kunst und Literatur thematisierte, von 2500 Menschen besucht worden waren.

> »A student reporter commented that »politics and art are easy to discuss and easy to understand. The impact of science is neither of these.« In all probability, the effective understanding that is sought in this program and many others will have to await extensive changes in the elementary and secondary schools through which the understanding of the scientific revolution will become an integral part of the total learning experience. The machine, however, may now be intervening in this affair.« (Ormsbee 1965, 163–164)

Die Maschine – damit waren die Computer gemeint und Robinson fuhr an gleicher Stelle so fort:

> »A graduate school dean recently observed that at his institution's computer center there had been a degree of contact and communication between the scientist and the humanists using the computer which far exceeds any contact between these groups which has developed in the past.« (Ormsbee 1965, 163–164)

Mehr als 30 Jahre später beobachtete Kevin Kelly, der Exekutivdirektor des Magazins »Wired« das internationale Aufkommen einer »The Third Culture«, die er in der Zeitschrift »Science« »Techno-Culture« nannte:

> »Techno-culture is not just an American phenomenon, either. The third culture is as international as science. As large numbers of the world's population move into the global middle class, they share the ingredients needed for the third culture: science in schools; access to cheap, hi-tech goods; media sat-

> uration; and most important, familiarity with other nerds and nerd culture. I've met Polish nerds, Indian nerds, Norwegian nerds, and Brazilian nerds. Not one of them would have thought of themselves as »scientists.« Yet each of them was actively engaged in the systematic discovery of our universe.« (Kelly 1998)

Eine »pop culture based in technology, for technology« sei entstanden, weil die Technologie unsere kulturelle Umwelt gesättigt habe, sie zu dominieren anstünde und es »cool« werde, ein »Nerd« zu sein, und es dränge sich die Frage auf, was dies für das Wissenschaftssystem und seine beiden Kulturen bedeutete:

> »The only reason to drag up this old rivalry between the two cultures is that recently something surprising happened: A third culture emerged. It's hard to pinpoint exactly when it happened, but it's clear that computers had a lot to do with it. What's not clear yet is what this new culture means to the original two.« (Kelly 1998)

Das Aufkommen der Computer hat das Wissenschaftssystem verändert, zunächst wandelte ihr Einsatz die experimentierenden Anteile der Natur- und Ingenieurwissenschaften, dann auch deren theoretischen Zweige aufgrund der großen Rechengeschwindigkeiten und Datenerfassungssysteme und schließlich nutzen die Forscher*innen auch die Textverarbeitungssysteme (Hashagen 2013). Bald profitierten auch die Geistes- und Sozialwissenschaften vom »computational turn« (Berry 2011) u.a. die Linguistik und die Statistik, und endlich setzte sich der »Computers als Werkzeug und Medium« (Friedewald 1999) im gesamten Wissenschaftssystem durch. Mit der Verbreitung des Internet und der Web-Dienste entstanden schließlich die Digital Humanities.

Die Geschichte der dritten Kultur als Techno-Kultur kann aber nicht auf die der Nutzung des Computers als Werkzeug reduziert werden. Sie wurde darüber hinaus von zwei Begriffen geprägt: Intelligenz und Information.

Energie, Intelligenz und Information

In der wissenschaftlichen Kultur der Natur- und Ingenieurwissenschaften waren Masse und Energie lange Zeit die beiden grundlegenden Größen. Zwar gibt es verschiedene Energieformen, diese lassen sich aber auf wissenschaftlich-

technischem Wege ineinander umwandeln und die Relativitätstheorie von Albert Einstein (1879–1955) zeigte, dass die Masse in Energie umwandelbar ist. Wenige Jahrzehnte später kam aber wieder eine neue Grundgröße hinzu: In seinem 1948 erschienenen Buch »Cybernetics: or Control and Communication in the Animal and the Machine« führte Norbert Wiener (1894–1964) die *information* als eine dritte solche Fundamentalgröße ein, die er auf die Gehirnaktivitäten zurückführte:

> »The mechanical brain does not secrete thought ›as the liver does bile,‹ as the earlier materialists claimed, nor does it put it out in the form of energy, as the muscle puts out its activity. Information is information, not matter or energy. No materialism which does not admit this can survive at the present day.« (Wiener 1948, 132)

Schon zum Ende der 1920er Jahre hatte der Physiker und Biologe Leó Szillard (1898–1964) der *information* den Status einer physikalischen Grundgröße zugesprochen. Seine Habilitationsarbeit handelt von »Eingriffen intelligenter Wesen« die bei einer Messung Information erhalten. Auch er bezog die Konzepte Intelligenz und Information aufeinander (Szilard 1929). Etwa gleichzeitig findet sich der Versuch eines Brückenschlags zwischen den Kulturen aus der amerikanischen Psychologie. Hier stellte David Wechsler (1896–1981) vergleichende Untersuchungen der menschlichen Fähigkeiten an, die damals als Maße für verschiedene menschliche Eigenschaften galten (Wechsler 1944). Die intellektuellen Fähigkeiten des Menschen ließen sich zwar quantitativ bewerten, indem die verschiedenen Aspekte dieser Fähigkeiten gemessen werden, aber sie seien nicht einfach zu einer Größe »Intelligenz« aufzusummieren. Intellektuelle Fähigkeiten und allgemeine Intelligenz seien nicht identisch und um dies zu untermauern bediente er sich einer Analogiebetrachtung zur Physik:

> »We do not, for example, identify electricity with our modes of measuring it. Our measurements of electricity consist of quantitative records of its chemical, thermal and magnetic effects. But these effects are not identical with the ›stuff‹ which produced them. General intelligence, like electricity, may be regarded as kind of energy. We do not know what the ultimate nature of this energy is, but as in the case of electricity, we know it by the things it does or, better, by the things it enables us to do—such as making appropriate associations between events drawing correct inferences from propositions, understanding the meaning of words, solving mathematical problems or building bridges. These are the effects of intelligence in the same sense as chemical

> dissociation, heat, and magnetic fields are the effects of electricity, but psychologists prefer the term mental products. We now intelligence by what it enables us to do.« (Wechsler 1944, 4)

Wechsler zufolge bewirkt die Intelligenz eines Wesens dessen Fähigkeiten. Auf diese verweisen messbare Größen, ähnlich wie die Elektrizität Eigenschaften der Materien bewirkt, auf die ebenfalls von messbaren Größen geschlossen wird. Diese Verknüpfung von Elektrizität und *intelligence* hat allerdings noch eine andere historische Wurzel, denn dem elektrischen Strom wurde noch vor einem Jahrhundert die Eigenschaft zugeschrieben, *intelligence* übertragen zu können.

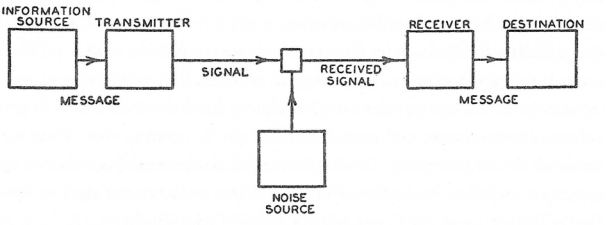
Schon seit dem 17. Jahrhundert wurde das telegraphisch zu Übertragende in der englischen (und auch in der französischen) Sprache mit dem Wort »intelligence« bezeichnet – ein Wort, das unter anderem die Fähigkeit kognitiv bzw. geistig etwas zu leisten, bedeutet, auch das geheime Einvernehmen oder Einverständnis zwischen Menschen. Im 18. Jahrhundert findet sich diese Bedeutung auch in den entsprechenden Übersetzungen in andere Sprachen: In Deutschland wurden amtliche Nachrichten in »Intelligenzblättern« mitgeteilt und »Intelligenz« war auch ein Name für die Gebildeten, im Russischen war das die »Intelligenzija«.

Das Wort *intelligence* stand auch für die Mitteilungen selbst: Schon 1664 sinnierte Robert Hook (1635–1703), der Kurator der Royal Society darüber, einen Apparat für »speedy intelligence« zu bauen, wenige Jahre später schlug er vor, »speedy conveyance of intelligence« von Ort zu Ort mit Hilfe von Teleskopen zu erreichen (Birch 1756, 299). Weitere 12 Jahre hielt er es für möglich, »to convey Intelligence from any one high and eminent Place, to any other that lies in sight of it« (Hook 1684). Auch der für seinen Telegraphier-Code bekannte Samuel Finley Breese Morse (1791–1872) gebrauchte das Wort »intelligence« in diesem Sinne: Als er im Herbst 1832 von den wissenschaftlichen Experimenten mit elektrischem Strom hörte und begriff, dass dessen Wirkung in jedem Teil eines stromdurchflossenen Drahtes sichtbar gemacht werden kann, schrieb er: »I see no reason why intelligence might not be instantaneously transmitted by electricity to any distance. » (Vail 1845, 70) Ein 1893 erschienenes Fachbuch des Elektroingenieurs Edwin James Houston (1847–1914) trug den Titel »The Electric Transmission of Intelligence: And Other Advanced Primers of Electricity« und am 9. Februar 1902 gründeten Albert Cushing Crehore (1868–1958) und George Owen Squier (1865–1934) die Crehore-Squier Intelligence Transmissi-

on Company, die mit ihrem landesweiten Telegraphensystem eine neue »art of transmitting intelligence« einzuführen versprach.

Heutige Begriffe, etwa »Daten« (*data*), »Nachricht« (*message*) und »Information« (*information*) waren in Wissenschaft und Technik noch nicht eindeutig definiert bzw. etabliert. Stattdessen wurde für alle drei das Wort *intelligence* benutzt, bis der Radio-Ingenieur Ralph Vinton Lyon Hartley (1888–1970) im September 1927 auf dem International Congress of Telegraphy and Telephony das Wort »information« in die Nachrichtentechnik einführte (Hartley 1928). Das Wort *intelligence* wurde seither in der Nachrichtenübertragung immer seltener benutzt, so dass seine Bedeutung als Nachricht oder Mitteilung bald verblasste.[4] Der Mathematiker und Begründer der »Mathematical Theory of Communication« Claude E. Shannon (1916–2001) gebrauchte nach dem II. Weltkrieg konsequent das Wort »information« für die transportierten Nachrichten zwischen Sender und Empfänger. In dem nach ihm benannte Kommunikationsschema (siehe Abb. 1) geht eine Nachricht von einer Informationsquelle zu einem Sender, dieser sendet deren Zeichen über den Kanal, der möglicherweise gestört wird, zum Empfänger, weshalb Shannon zwischen dem Zeichen vor dem Kanal und dem empfangenen Zeichen hinter dem Kanal unterschied. Vom Empfänger geht dann eine Nachricht zum Kommunikationsziel.

Abb. 1: Das Kommunikationsschema von Shannon (Shannon 1948, 381).



4 In der englischen Sprache erinnern einige Worte daran: Die »Central Intelligence Agency« (CIA) hat als US-amerikanischer Auslandsgeheimdienst die Aufgabe, vor allem Information über Menschen durch Menschen zu beschaffen, im Gegensatz zur Beschaffung mit technischen Mitteln, die als Signals Intelligence bezeichnet wird. Auch der deutsche Geheimdienst BND verwendet entsprechende Bezeichnungen und Abkürzungen, siehe: https://www.bnd.bund.de/DE/Die_Arbeit/Informationsgewinnung/informationsgewinnung_node.html.

Die mathematische Theorie der Kommunikation, etablierte sich in den 1950er Jahren innerhalb der Elektrotechnik bei deren Ausdifferenzierung in verschiedene Spezialgebiete unter dem Namen »Information Theory«. *Information* war darin eine technisch messbare Größe. Ihre Bedeutung und die mit ihr verbundenen Absicht bzw. Wirkung, also der Grund für ihre Sendung und was durch sie erreicht wurde, waren nicht Gegenstand der Theorie, und entsprechende spätere Anwendungen seiner Theorie in den Geistes- und Sozialwissenschaften fand Shannon suspekt! (Tribus 1978)

Genau diese Aspekte sollten aber auch zum Begriff *information* gehören, forderte der Mathematiker und Wissenschaftsorganisator Warren Weaver (1894–1978) in einem Artikel für den *Scientific American*, der im Juli 1949 erschien (Weaver 1949). Er unterschied darin drei Problemebenen dieses Begriffs:

- Ebene A betrifft die syntaktischen Eigenschaften.
- Ebene B betrifft die semantischen Eigenschaften.
- Ebene C betrifft die pragmatischen Eigenschaften.

Für Weavers weiteren Informationsbegriffs interessierten sich auch Wissenschaftler*innen der geistes- und sozialwissenschaftlichen Kultur. Für den Psychologen und Philosophen Gerhard Maletzke war die Informationsübertragung die Beziehung, die Lebewesen untereinander eingehen können: Sie verständigen sich, sie sind imstande, innere Vorgänge oder Zustände auszudrücken, sie teilen ihren Mitgeschöpfen Sachverhalte mit oder sie fordern sie zu einem bestimmten Verhalten auf. Pointiert schrieb er: »Kommunikation ist die Bedeutungsvermittlung zwischen Lebewesen.« (Maletzke 1963, 16)

Artificial Intelligence I

Das interdisziplinäre Interesse an der Informationstheorie war groß. Zur Psychologie wie auch zur Hirnforschung wurden Schnittstellen gesucht. Umgekehrt interessierten sich Mathematiker*innen, Nachrichtentechniker*innen und Computerkonstrukteur*innen für das logische Modell von Nervenzellen und deren Vernetzung, das Warren Sturgis McCulloch (1898–1969) und Walter Pitts (1923–1969) 1943 publiziert hatten (McCulloch and Pitts 1943). Auf dem Hixon Symposium über »Cerebral Mechanisms in Behavior«, das 1948 am California Institute of Technology in Pasadena stattfand (Jeffress 1951), hielt Mc-

Culloch den Vortrag »Why the Mind Is in the Head« (McCulloch 1951) und John von Neumann (1903–1957) sprach über »The General and Logical Theory of Automata« (von Neumann 1951). Auf diese Forscher sowie u.a. auch auf Shannon traf hier der junge Mathematiker John McCarthy (1927–2011), der seine Teilnahme am Hixon-Symposium später sein »watershedmoment« nannte (Hayes 2007):

> »At this symposium, the computer and the brain were compared, the comparison being rather theoretical, since there wern't [sic!] any stored programmed computers yet. The idea of intelligent computer programs isn't in the proceedings, though maybe it was discussed. Turing already had the idea in 1947. I developed some ideas about intelligent finite automata but found them unsatisfactory and didn't publish.« (McCarthy 2002)

Vier Jahre später war McCarthy zu einem halbjährigen Forschungsaufenthalt in den Bell Laboratorien, wo er mit Shannon die Möglichkeiten »intelligente Maschinen« zu konstruieren diskutierte. Die beiden beschlossen, einen Sammelband zu diesem Thema herauszugeben. Shannon fand den Begriff »Machine Intelligence« allerdings »much too flashy«, und so einigten sie sich für das im Jahre 1956 erschienene Buch auf den Titel »Automata Studies« (Shannon and McCarthy 1956). Dass die meisten darin publizierten Arbeiten nicht den Intelligenzbegriff, sondern mathematisch-logische Automaten thematisierten, war für McCarthy allerdings enttäuschend.

In diesen Jahren freundete sich McCarthy mit dem Mathematiker Marvin Lee Minsky (1927–2016) an, der damals Junior Fellow in Mathematik und Neurologie in Harvard war. Mit Hilfe des Physik-Doktoranden Dean Stockett Edmonds (1924–2018) hatte Minsky 1951 an der University in Princeton aus Vakuumröhren und einem Motor einen »neuronalen Netzwerksimulator« gebaut: SNARC (Stochastic Neural-Analogue Reinforcement Computer) wurde die Grundlage für seine Dissertation (Minsky 1954). Es war diese gemeinsame Leidenschaft für die Konstruktion intelligenter Maschinen, die McCarthy und Minsky zu zwei der wichtigsten Pioniere der KI werden ließ.

McCarthy wurde 1955 Assistenzprofessor am Dartmouth College in Hanover, New Hampshire. Dort lernte er den IBM-Elektroingenieur Nathaniel Rochester (1919–2001) kennen. Mit Minsky, Shannon und Rochester verfasste er im August 1955 »A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence« (McCarthy, Shannon, Minsky, and Rochester 1955). Den

Begriff »Artificial Intelligence«, der zuvor nicht gebräuchlich war, führten sie folgendermaßen ein:

> »The study is to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it.« (McCarthy, Shannon, Minsky, and Rochester 1955)

Unter Artificial Intelligence verstanden die Antragsteller die Simulation der höheren Funktionen des menschlichen Gehirns, die ihm/ihr jene Fähigkeiten und Verhaltensweisen, ermöglichen, die »intelligent« genannt werden. Sie nannten sieben Forschungsschwerpunkte: Für die Grundlage solcher Simulationen waren (1) »Automatic Computers« notwendig, dann war zu fragen (2) »How Can a Computer be Programmed to Use a Language«, was auf damalige Spekulationen verweist, dass menschliches Denken auf Wortmanipulationen entsprechend logischen Regeln basiert. Zudem sollten (3) »Neuron Nets« erforscht werden, wobei damalige Theorien besagten, dass gewisse Neuronenverbindungen zu Begriffsbildungen führen. Gesucht wurde (4) eine »Theory of the Size of a Calculation«, die Effizienzüberlegungen berücksichtig, man ging davon aus, dass (5) eine »intelligente« Maschine zu einem gewissen »Self-lmprovement« und (6) zu »Abstractions« von Sensorwerten und anderen Daten fähig ist. Zuletzt wurden (7) »Randomness and Creativity« angeführt, da die Berücksichtigung von Zufälligkeiten den Unterschied zwischen kreativem Denken und phantasielosem, kompetentem Denken ausmache (McCarthy, Shannon, Minsky, and Rochester 1955).

Die Rockefeller Foundation genehmigte das Projekt und so konnte vom 19. Juni bis zum 16. August 1956 dieses als »Geburtsstunde der Artificial Intelligence« in die Wissenschaftsgeschichte eingegangene Treffen stattfinden. Weder wurden hier aber tatsächlich Versuche unternommen, Maschinen Sprachen benutzen zu lassen, noch sie Probleme lösen zu lassen, die bis dahin dem Menschen zu lösen vorbehalten waren.

Herbert Alexander Simon (1916–2001), Alan Newell (1927–1992) und John Clifford Shaw (1922–1991) hatten an Programmen zur Lösung von »ultracomplicated« Problemen, etwa aus den Bereichen des Schachspielens, der Euklidischen Geometrie, der Streichholzaufgaben oder der symbolischen Logik gearbeitet. Bei der Rand Corporation in Santa Monica in Kalifornien hatten sie den »Logic Theorist« entworfen – ein System, das Beweise einiger Theoreme durchführte, wie sie in der »Principia Mathematica« vorexerziert wurden. Die-

ses System stellten sie bei dem Dartmouth-Treffen vor und auch ein noch nicht vollständig fertiges Schachprogramm von Alex Bernstein (1930–1999) wurde dort diskutiert (Bernstein 1958).

Danach beschäftigten sich einige der AI-Pioniere eingehend mit Datenlistenstrukturen. Minsky hatte noch dort mit Rochester ausführlich die Möglichkeiten erörtert, ähnlich dem Logic Theorist, der Theoreme der Aussagenlogik bewies, auch ein Beweisprogramm für geometrische Sätze zu konstruieren. Rochester besprach dies mit seinem neuen Mitarbeiter Herbert Leo Gelernter (1929–2015), der noch im gleichen Sommer gemeinsam mit seinem Kollegen Carl L. Gerberich ein solches Programm schrieb. Auf Anraten von McCarthy erweiterten sie die Sprache Fortran durch einige Listenoperationen zu FLPL (Fortran List Processing Language). McCarthy selbst publizierte dann zwei Jahre später mit Hilfe von Rochester, der 1958 Visiting Professor am MIT war, LISP als eine mächtigere »List Processing«-Sprache. Schließlich ist das Programm SAINT (Symbolic Automatic INTegrator) zu nennen, das Minskys Mitarbeiter James Robert Slagle (1934–1994) im Jahre 1961 in seiner Dissertation zur Lösung von Analysis-Aufgaben entwarf (Slagle 1963).

Artificial Intelligence II

Schon im Jahre 1950 hatte der englische Mathematiker Alan Mathison Turing (1912–1954) in der psychologischen Zeitschrift »Mind« seinen Artikel »Computing Machinery and Intelligence« publiziert (Turing 1950). Er hatte das »Imitationsspiel« eingeführt, das zur Klärung der Frage »Can machines think?« beitragen sollte. Diese Arbeit hatten McCarthy und Shannon auch schon im Vorwort zu den »Automata Studies« erwähnt:

> »The problem of giving a precise definition to the concept of ›thinking‹ and of deciding whether or not a given machine is capable of thinking has aroused a great deal of heated discussion. One interesting definition has been proposed by A. M. Turing: a machine is termed capable of thinking if it can, under certain prescribed conditions, imitate a human being by answering questions sufficiently well to deceive a human questioner for a reasonable period of time. A definition of this type has the advantages of being operational or, in the psychologists' term, behavioristic. No metaphysical notion of consciousness, ego and the like are involved. While certainly no machines

> at the present time can even make a start at satisfying this rather strong criterion, Turing has speculated that within a few decades it will be possible to program general purpose computers in such a way as to satisfy this test.« (Shannon and McCarthy 1956, V)

Es gibt nur einen Beitrag in den »Automata Studies«, dessen Überschrift das Wort »intelligence« enthält, und der damit das Vermögen kognitiver Fähigkeiten bezeichnete, die auf seine Gehirnleistung zurückzuführen sind: »Design for an Intelligence-Amplifier« vom britischen Psychiater William Ross Ashby (1903–1972). Dieser verarbeitete darin auch psychologische Arbeiten zur Intelligenz aus der ersten Hälfte des 20. Jahrhunderts, darunter Wechslers Buch »Measurement of Adult Intelligence« (Wechsler 1944) und »Measuring Intelligence« (Terman and Merrill 1937) des amerikanischen Psychologen Lewis Madison Terman (1877–1956) und seiner Mitarbeiterin Maud Amanda Merrill (1888–1978), die darin das Konzept des Intelligenzquotienten von William Louis Stern (1837–1890) übernommen hatten. Ashby zog es heran, um zu argumentieren, dass die menschliche Intelligenz begrenzt ist und ihre Verstärkung durch Maschinenleistungen zur Lösung der vielen anstehenden und sehr komplexen Probleme von größtem Nutzen sei. Computer seien dazu geeignet, dass, ähnlich wie mit Hilfe von Kraftmaschinen – Ashby nannte sie »power-amplifier« – die menschliche Fähigkeit Kraft auszuüben bzw. Arbeit zu verrichten verstärkt wird. Entsprechend seien die Computer »intelligence-amplifier«, die also keine eigene Intelligenz haben, sondern »synthetischen intellektuellen Fähigkeiten«, mit deren Hilfe die menschliche Intelligenz verstärkt werden kann. Damit hat Ashby eine Unterscheidung zwischen natürlicher und einer artificial intelligence angedeutet, die allerdings in dem kurz darauf entstandenen Forschungsgebiet Artificial Intelligence keine Berücksichtigung fand. Für ihn bedeutete menschliche Intelligenz die natürlichen kognitiven Fähigkeiten, während die maschinelle oder »artificial« Intelligenz für ihn eine nachrichtentechnische Bedeutung hatte, mit deren Hilfe die natürliche Intelligenz zwar verstärkt, aber eben *nur* verstärkt werden könne. Hingegen hatten die Antragsteller für das Dartmouth-Treffen bei der Attribuierung der *intelligence* als »artificial« den damals ambivalenten Gebrauch des Wortes *intelligence* nicht thematisiert.

Von der Statistik über die Datenwissenschaft zum Maschinellen Lernen

Zur Mitte der 1980er Jahre prägte der Philosoph John Haugeland (1945–2010) für die Artificial Intelligence, wie sie bisher beschrieben wurde, die Bezeichnung GOFAI (Good Old Fashioned Artificial Intelligence). Er grenzte sie von neueren AI-Ansätzen ab, die sich künstlicher neuronaler Netze und Klassifikationsbäumen bedienten (Haugeland 1985). GOFAI ging von der Annahme aus, dass Aspekte der Intelligenz durch die Manipulation von Symbolen in einer Maschine erreicht werden können. Dazu wurden Algorithmen programmiersprachlich formuliert und dann Befehle nach Befehl abgearbeitet.

Künstliche neuronale Netze und Klassifikationsbäume sind dagegen Algorithmen, die Muster in Datenmengen suchen. Sie bewerkstelligt auf diese Weise etwas, das Menschen schlecht bzw. gar nicht können, nämlich große Datenmengen zu beherrschen. Insofern können sie »Intelligence Amplifier« sein, weil sie komplementär zur menschlichen Intelligenz Resultate erzielen und so deren Auswirkungen verstärken können. Schon 1959 hatte der Elektroingenieur Arthur Lee Samuel (1901–1990) überlegt, wie Computer befähigt werden können, zu lernen, ohne dass sie dafür explizit programmiert werden (Samuel 1959). Sein Programmcode für das Brettspiel »Dame« enthielt lediglich die Regeln des Damespiels. Die Güte des Damespiels des Programms wurde aufgrund jener Daten gesteigert, die das Programm erhält, wenn bestimmte Spielkonstellationen, die vorher nicht erreicht worden waren, eine bessere Bewertung bekamen. So konnte das Damespielen des Programms optimiert werden, während das Programm selbst unverändert blieb.

Die Ergebnisse wurden mit Wahrscheinlichkeiten bewertet, somit sind die entsprechenden Algorithmen probabilistisch, denn anders als die klassischen Algorithmen liefern sie Ergebnisse, die nur mit einer gewissen Wahrscheinlichkeit richtig sind, während die klassischen Algorithmen stets die eindeutige Lösung des Problems finden. Die so programmierten Algorithmen »lernen« in dem Sinne, dass ihr Output mit der Zeit verbessert werden kann. Die Programme führen also nicht bei jedem Durchlauf auf denselben Ausgabewert. Bei der Bewältigung ihrer eng begrenzten Aufgaben können ML-Algorithmen sich somit immer weiter verbessern.

Große Expertise im Umgang mit Daten hatten lange vor den Computern die Statistiker*innen erworben, die schon in ihren frühen Anfängen als »Staatsbeschreibung« vor allem Daten erhoben, gesammelt und analysiert hatte. Die Statistik wurde zu einer Wissenschaft, die das Verständnis von

Daten fördert und daran anschließend Hypothesen zu generieren imstande ist und ihre Methoden wurden in allen empirisch arbeitenden Wissenschaften wichtig. Ihre Geschichte verläuft zwischen mathematischer Theorie, wissenschaftlichem Berechnen und Anwendungen. Mit letzteren begann ihre historische Entwicklung, die aber bald einer Mathematisierung unterzogen wurde. Das mathematisch-statistische Methodengebäude entstand bis zur Mitte des 20. Jahrhunderts nahezu ungestört, bis im Jahre 1950 die Arbeit »Statistical Decision Functions« von Abraham Wald (1902–1950) erschien (Wald 1949). Sie steht für einen gewissen Abschluss der Mathematisierung der Statistischen Inferenz, wie die beiden Statistiker und Stanford-Professoren Bradley Efron (*1938) und Trevor Hastie (*1953) im Epilog ihres vor acht Jahren erschienenen Buchs »Computer Age Statistical Inference« schreiben (Efron and Hastie 2016). Bis dahin spielte der empirisch-numerische Zugang, ihrer Ansicht nach kaum eine Rolle. Erst mit Aufkommen der elektronischen Computer veränderte sich die Statistik-Entwicklung in der zweiten Hälfte des 20. Jahrhunderts und es setzte ein Prozess ein, der die Statistik aus ihrem »Eigenbrötlertum« um mathematische Strukturen herausgelöst und mitgerissen habe, argumentieren die beiden Autoren des Buchs, das den Untertitel »Algorithms, Evidence, and Data Science« bekam.

Bereits 1962 hatte der Statistiker John Wilder Tukey (1915–2000) von seinen Kollegen im Artikel »The Future of Data Analysis« eine anwendungs- und berechnungs-orientierte Ausrichtung der Statistik als Disziplin gefordert (Tukey 1962). Er beurteilte den zur erwartenden »impact of the computer« realistisch, wenn er schrieb, dass die Datenanalyse »could be done by hand on small data sets, but […] speed and economy of delivery of answer make the computer essential for large data sets and very valuable for small sets.« Und: »The future of data analysis can involve great progress, the overcoming of real difficulties, and the provision of a great service to all fields in science and technology.« (Tukey 1962, 64) Acht Jahre später wies er den Statistikern den Weg von den damals dominanten mathematischen Methoden der statistischen Analyse zur datengenerierten Herleitung von Hypothesen. Gemeinsam mit Jerome Harold Friedman (*1939) entwickelte er die statistische Methode »Projection Pursuit zur Problemlösung bei Daten mit sehr viele Dimensionen. Bei dem Verfahren wird in den hochdimensionalen Datenraum eine Hyperebene gelegt, auf die die Daten zu projiziert werden. In diesen Projektionen lassen sich »interessante« Strukturen aufdecken (Friedman and Tukey 1974).

Im Juli 1982 prophezeite Tukey in seinem Vortrag »Another Look at the Future« für das 14. Interface-Symposium, dass die künftigen Computer die

Datenanalyse gravierend verändern würden: »data analysis would become so computationally intensive that it would push the limits of existing computer systems.« Deswegen plädierte er für interdisziplinäre Zusammenarbeit der Kulturen der Statistik und der Computer Science:

> »This means (i) large systems, (ii) systems planned both for growth and for easy specialized attachment, (iii) cooperation between a variety of insightful data analysts on the one hand and a variety of computer experts on the other – each group with diverse skills. Success will not be easy, but starting now poses no major barriers. There are people with enough insights of the needed kinds, though they may be hard to find and assemble. And we can expect the 4th or 5th generations of such systems to be far, far better than anything we have today.« (Tukey 1982)
>
> »Something important changed in the world of statistics in the new millennium« schrieben Efron und Hastie, und die Ursache dafür sahen sie in den Möglichkeiten der Voraussagealgorithmen die das Fach Statistik zu »Data Analytics« und schließlich zur »Data Science« wandelten. Mit ihren Erfolgen, die sich mit Big Data einstellten, wurden sie immer wichtiger. Efron und Hastie nennen sie die »media stars of the Big-Data era« (Efron and Hastie 2016, 446).

Im Jahre 2001 publizierte Leo Breimans (1928–2005) den Artikel »Statistical Modeling: The Two Cultures«, der im gleichen Heft von zahlreichen Fachkolleg*innen kommentiert wurde (Breiman 2001). Innerhalb der Statistik gebe es zwei unterschiedliche »Kulturen« des statistischen Modellierens: Von den Daten zu ihren Schlussfolgerungen gelangten Statistiker*innen entweder ausgehend von der Prämisse, dass ein gegebenes stochastisches Datenmodell die Daten erzeugt oder sie verwenden algorithmische Modelle ohne Vorannahmen über einen Datenmechanismus. Die traditionelle Statistik-Kultur beruhe auf dem Glauben, dass ein(e) Statistiker*in durch Vorstellungskraft und Blick auf die Daten eine einigermaßen gute parametrische Klasse von Modellen für das Naturgeschehen entwickeln kann, und daraufhin würde sie/er die Parameter schätzen und Schlussfolgerungen ziehen. Breiman hielt diese Modelle nicht mehr für genügend, weil die betrachteten Systeme ungeheuer groß und komplex sind und aus den Wissenschaften immer mehr Fragen aufkamen. So mussten auch die Datenstrukturen immer komplexer werden und es wurde schwieriger, geeignete Datenmodelle zu konstruieren.

In der »industriellen Statistik« hatte sich seit Jahrzehnten ein alternatives Vorgehen entwickelt und etabliert. Hochkomplexe Probleme, die

sich der Datenmodellierung entzogen, fanden sich z.B. vermehrt bei der Sprach- und Bilderkennung bzw. -verarbeitung, sowie bei der Voraussage von nichtlinearen Zeitreihen und Finanzmarktanalysen. Seit der Mitte der 1980er Jahre setzten in diesen Feldern vor allem jüngere Informatiker*innen, Physiker*innen und Ingenieure*innen auf die neuen algorithmischen Modelle der künstlichen neuronalen Netzwerke oder der Entscheidungsbäume. Auch Psychometriker*innen und Sozialwissenschaftler*innen nutzten diese Verfahren. Zu dieser neuen Wissenschaftskultur der algorithmischen Modellierung gehörten damals aber kaum Statistiker*innen aus den Universitäten:

> »[…] the list of statisticians in the algorithmic modeling business is short, and applications to data are seldom seen in the journals. The development of algorithmic methods was taken up by a community outside statistics.« (Breiman 2001, 205)

Die Kultur der algorithmischen Modellierung nutzt Algorithmen zur Klassifizierung von Daten und zu ihrer Voraussage aufgrund schon vorliegender Daten, und wegen ihrer hohen Prognosegenauigkeit war sie sehr erfolgreich. Sie basierte auf zwei wichtigen Veränderungen: »[to] challenge for the tools and computers of the time« (Cutler 2010, 1622) und »[to] make the transition from probability theory to algorithms« (Breiman 2001, 215).

Friedman hatte diese »Data-Mining-Revolution« erwartet, und auch gesehen, dass sie das Fach Statistik an einen Scheideweg bringt. Er empfahl seiner Zunft: »make peace with computing« und »moderate our romance with mathematics« (Friedman 1997, 6). Im gleichen Jahr forderte Chien-Fu Jeff Wu (*1949) die Statistiker*innen-Community auf das Fach Statistik in »Data Science« umzubenennen und nicht mehr von »Statistiker*innen«, sondern von »Data scientists« zu sprechen: »It is time in the history of statistics to make a bold move« (Wu 1997, 11). Man möge sich auf die »großen Datenmengen« fokussieren, sich den anderen Wissenschaften mehr öffnen – auch für die Ausbildung von Datenwissenschaftler*innen sollten die anderen Wissenschaften treibend sein –, deren empirisch-physikalischen Ansatz und deren Wissen zur Problemlösung nutzen.

Zum Ende des 20. Jahrhunderts waren für die statistische Inferenz computerintensive Algorithmen eingeführt worden,[5] die Software-Packages SAS

5 Die israelischen Statistiker Yoav Benjamini (*1949) und Yosef (Yosi) Hochberg (1945–2013) publizierten die Falscherkennungsrate (englisch False Discovery Rate), die

und SPSS waren weit verbreitet, es entstanden Mathematica und Matlab, und die auf statistischen Computersprachen S und ihrem »Open Source«-Nachfolger R aufsetzenden Algorithmen kamen hinzu (Ross and Gentleman 1996). Zudem wurden die Algorithmen der künstlichen neuronalen Netze, und des Deep Learning sowie andere Machine-Learning-Algorithmen populär. Auch der Statistiker William Swain Cleveland (*1943) warb nun dafür, dass aus der Statistik eine Data Science hervorgehen möge, indem sich die Statistiker*innen den technischen Aspekten der Disziplin der Informatik zuwenden und mit deren Vertreter*innen zusammenzuarbeiten (Cleveland 2001).

Noch im Jahre 1997 hatte Friedman in einem Keynote-Vortrag auf dem 29. Symposium on the Interface Between Computer Science and Statistics davor gewarnt, dass die Statistiker*innen, die auf sie zukommende Data-Mining-Revolution verpassen könnten (Friedman 1997). Gut ein Jahrzehnt später erklärte Wilkinson diese Revolution für beendet: »we are in an era of *machine learning*« (Wilkinson 2008, 419).[6]

Schluss

Der Begriff der *intelligence* ist in den letzten Jahrhunderten in verschiedenen Wissenschaftskulturen und in verschiedenen Bedeutungen benutzt worden. In der Nachrichtentechnik stand *intelligence* seit Beginn der Überlegungen zur Telegraphie bis ins erste Drittel des 20. Jahrhundert für die Mitteilungen und damit sowohl für die übertragenen Nachrichten als auch für deren Bedeutung. In der Psychologie wurde mit *intelligence* die Ursache für die intellektuellen Eigenschaften der Menschen bezeichnet, später auch anderer Lebewesen, und mit Aufkommen der Computer schon bald auch der Maschinen. Schon die frühen Versuche diese verschiedenen Begrifflichkeiten zu klären und so die später von Snow benannte Kluft zwischen diesen Wissenschaftskulturen zumindest an einer Stelle zu überbrücken, stifteten Verwirrung, die noch verstärkt wurde, als der *intelligence* das Attribut *artificial* mitgegeben wurde. Damit war die in Maschinen programmierte Simulation der intellektuellen Eigenschaften

ein Testverfahren zur Beherrschung multipler Testprobleme liefert und Robert Tibshirani (*1956) formulierte die »Lasso« (Least absolute shrinkage and selection operator) genannte Regressionsanalysemethode. Siehe dazu: (Tibshirani 1996).

6 Kursive im Original.

von Menschen gemeint, doch auch die hypothetische Annahme, dass Computer bzw. Computerprogrammen selbst diese intellektuellen Eigenschaften zugeschrieben werden könne und sie deshalb intelligent zu nennen seien, wurde vertreten. Und dann gab es auch noch die Interpretation der Artificial Intelligence als computerisierte Verstärkung der menschlichen Intelligenzleistungen. In den letzten Jahrzehnten wird unter Artificial Intelligence nun das Machine Learning verstanden, das von datengetriebenen Algorithmen handelt, die gerade solche Leistungen hervorbringen, zu denen Menschen nicht in der Lage sind. Für den Fall, dass jemand den technischen Dingen überhaupt *intelligence* zuschreiben will, könnte das Akronym AI dann zu »Alternative Intelligence« aufgelöst werden.

Literatur

Adorno, Theodor W. 1966. *Negative Dialektik*. Frankfurt a.M.: Suhrkamp.

Bernstein, Alex. 1958. »A Chess Playing Program for the IBM 704.« *Chess Review* 26 (7): 208–209.

Berry, Dave M. 2011. »The Computational Turn: Thinking About the Digital Humanities.« *Culture Machine*, 12. Accessed July 4, 2023. https://sro.sussex.ac.uk/id/eprint/49813/1/BERRY_2011-THE_COMPUTATIONAL_TURN-_THINKING_ABOUT_THE_DIGITAL_HUMANITIES.pdf.

Bibel, Wolfgang, and Ulrich Furbach. 2018. *Formierung eines Forschungsgebiets – Künstliche Intelligenz und Intellektik an der Technischen Universität München*. Deutsches Museum Preprint 15.

Birch, Thomas. 1756. *The History of the Royal Society of London for Improving of Natural Knowledge From Its First Rise In Which The most considerable of those Papers communicated to the Society, which have hitherto not been published, are inserted in their proper order, As A Supplement To The Philosophical Transactions, IV*. London: A. Millar in the Strand.

Blasche, Siegfried. 2008. »Geist, objektiver«. In: *Enzyklopädie Philosophie und Wissenschaftstheorie*, edited by Jürgen Mittelstraß. Stuttgart: Metzler, Band 4, 722–724.

Brockman, John. 1995. *The Third Culture: Beyond the Scientific Revolution*. New York: Simon & Schuster.

Carnap, Rudolf. 1928. *Der logische Aufbau der Welt*. Berlin-Schlachtensee.

Cleveland, William S. 2001. »Data Science: an Action Plan for Expanding the Technical Areas of the Field of Statistics.« *International Statistical Review* 69 (1): 21–26.

Cleveland, William S. 2019. *The Collected Works of John W. Tukey. V: Graphics 1965–1985*. Boca Raton, FL: Chapman and Hall/CRC.

Dilthey, Wilhelm 1883. *Einleitung in die Geisteswissenschaften, Versuch einer Grundlegung für das Studium der Gesellschaft und ihrer Geschichte*. Leipzig: Dunckler & Humblot.

Dilthey, Wilhelm 1910. *Der Aufbau der geschichtlichen Welt in den Geisteswissenschaften*. Berlin: Abhandlungen der preußischen Akademie der Wissenschaften, Philosophisch-Historische Klasse, Jg. 1910, 1–123.

Efron, Bradley, and Trevor Hastie. 2016. *Computer Age Statistical Inference. Algorithms, Evidence, and Data Science*. Cambridge University Press.

Friedewald, Michael. 1999. *Der Computer als Werkzeug und Medium. Die geistigen und rechnerischen Wurzeln des Personal Computers*. Berlin-Diepholz: GNT-Verlag.

Friedman, Jerome H. 1997. *Data Mining and Statistics: What's the Connection?* Keynote address in Computing Science and Statistics: Proceedings of the 29th Symposium on the Interface Between Computer Science and Statistics.

Friedman, Jerome H., and John W. Tukey. 1974. »A Projection Pursuit Algorithm for Exploratory Data Analysis.« *IEEE Transactions on Computers C-23* (9): 881–890.

Hartley, Ralph Vinton Lyon. 1928. »Transmission of Information.« *The Bell System Technical Journal VII* (3): 535–563.

Hashagen, Ulf. 2013. »The Computation of Nature, Or: Does the Computer Drive Science and Technology?« In: *The Nature of Computation. Logic, Algorithms, Applications (CiE 2013)*, edited by Paola Bonnizzoni, Vasco Brattka, and Benedikt Löwe. Berlin, Heidelberg: Springer, 263–270.

Haugeland, John. 1985. *Artificial Intelligence: The Very Idea*. Cambridge, Mass.: MIT Press.

Hayes, Patrick J., and Leora Morgenstern L. 2007. »On John McCarthy's 80th Birthday, in Honor of His Contributions« *AI Magazine* (Winter): 93–102.

Hook, Robert. 1684. *Philosophical experiments and observations of the late eminent Dr. Robert Hooke, and Geom. Prof. Gresh, and other eminent virtuoso's in his time, Discourset of the Royal Society, for 21. Mai 1684*. London: William Derham, 1635–1703.

Jeffress, Lloyd A. (ed.). 1951. *Cerebral Mechanisms in Behavior: the Hixon Symposium*. New York: Wiley.

Jordan, Pascual. 1932. »Die Quantenmechanik und die Grundprobleme der Biologie und Psychologie.« *Die Naturwissenschaften* 45: 815–821.

Kay, Lily E. 2002. *Das Buch des Lebens. Wer schrieb den genetischen Code?* München: Hanser.

Kelly, Kevin. 1998. »The Third Culture.« *Science* 279 (5353): 992–993.

Maletzke, Gerhard. 1998. *Psychologie der Massenkommunikation*. Hamburg: Verlag Hans Bredow-Institut.

Maletzke, Gerhard. 1998: *Kommunikationswissenschaft im Überblick*. Opladen/Wiesbaden: Westdeutscher Verlag.

McCarthy, John. 2006. *Dartmouth and Beyond, 2006*. Accessed July 4, 2023. http://www-formal.stanford.edu/jmc/slides/dartmouth/dartmouth-sli/.

McCarthy, John et al. 1955. *A proposal for the Dartmouth summer research project on artificial intelligence, August 31, 1955*. Accessed July 4, 2023. www-formal.stanford.edu/jmc/history/dartmouth/dartmouth.html.

McCulloch, Warren S. 1951. »Why the Mind is in the Head.« In: *Cerebral Mechanisms in Behavior: the Hixon Symposium*, edited by Lloyd A. Jeffress. 42–111. New York: Wiley.

McCulloch, Walter S. and Walter Pitts. 1943. »A logical calculus of the ideas immanent in nervous activity« *Bulletin of Mathematical Biophysics* 5: 115–133.

Ormsbee, W. Robinson 1965. Review of: C. P. Snow: The Two Cultures: And a Second Look. *Technology and Culture 6 (1) Museums of Technology* (Winter): 162–164.

Minsky, Marvin Lee. 1952. *Neural-Analogue Calculator Based upon a Probability Model of Reinforcement*. Harvard University Psychological Laboratories: Cambridge, Mass., January 8.

Minsky, Marvin Lee. 1954. »Theory of neural-analog reinforcement systems and its application to the brain-model problem.« PhD diss., Princeton University, N.J.

Ross, Ihaka, and Robert Gentleman. 1996.: »R: A Language for Data Analysis and Graphics.« In: *Journal of Computational and Graphical Statistics* 5 (3). American Statistical Association, Institute of Mathematical Statistics, Interface Foundation of North America, Alexandria: 299–314. Accessed July 4, 2023. https://www.stat.auckland.ac.nz/~ihaka/downloads/R-paper.pdf.

Samuel, Arthur Lee. 1959. »Some Studies in Machine Learning Using the Game of Checkers.« *IBM Journal of Research and Development* 3 (3): 210–229.

Shannon, Claude E. 1948. »The Mathematical Theory of Communication.« *Bell System Technical Journal* 27 (3, 4): 379–423, 623–656.

Shannon, Claude E., and John McCarthy, ed. 1956. *Automata Studies*. Princeton, NJ: Princeton University Press (Annals of Mathematical Studies).

Shermer, Michael. 2007. »The Really Hard Science.« *Scientific American* 297 (4): 44–46.

Slagle, James R. 1963. »A Heuristic Program That Solves Symbolic Integration Problems in Freshman Calculus, Symbolic Automatic Integrator.« *Journal of the ACM* 10 (4): 507–520.

Snow, Charles Percy. 1959. *The Two Cultures and the Scientific Revolution (The Rede Lecture 1959)*. Cambridge, Mass.

Snow, Charles Percy. 1964. *Two Cultures: And a Second Look. An Expanded Version of the Two Cultures and the Scientific Revolution*. Cambridge, Mass.

Szilárd, Leo. 1929. »Über die Entropieverminderung in einem thermodynamischen System bei Eingriffen intelligenter Wesen.« *Zeitschrift für Physik* 53: 840–856.

Terman, Lewis M., and Maud A. Merrill. 1937. *Measuring Intelligence: A Guide to the Administration of the New Revised Stanford-Binet Tests of Intelligence*. London, George G. Harrap.

Tibshirani, Robert. 1996. »Regression shrinkage and selection via the lasso.« *Journal of the Royal Statistical Society, Series B* 58 (1): 267–288.

Tribus, Myron T. 1978. »Thirty Years of Information Theory.« In: *The Maximum Entropy Formalism*, edited by Levine Raphael D., Tribus, Myron T. The MIT Press, Cambridge MA, 1–14.

Tukey, John W. 1962. »The Future of Data Analysis.« *The Annals of Mathematical Statistics* 33 (1): 1–67.

Tukey, John W. 1982. »Another look in the future.« In: *Computer Science and Statistics, Proceedings of the 14th Symposium of the Interface*, edited by Heiner, Karl W., Sacher, Richard S. Wilkinson, John W. New York: Springer: 2–8.

Turing, Alan M. 1950. »Computing Machinery and Intelligence.« *Mind* 49 (236): 433–460.

Ungeheuer, Gerold, Dieter Krallmann, Helmut O. Schnelle, and Hans G. Tillmann. 1966. »Künstliche Intelligenz«, Forschungsbericht 66–7, Gutachterauftrag T-596-L-203, Institut für Phonetik und Kommunikationsforschung, Universität Bonn.

Vail, Alfred. 1845. *The American Electric Magnetic Telegraph: With the Reports of Congress and a Description of All Telegraphs Known, Employing Electricity or Galvanism*. Philadelphia: Lea and Blanchard: S. 70.

Wechsler, David. 1944. *The Measurement of Adult Intelligence*. Third Edition, Baltimore: The Williams & Wilkins Company (1. ed. 1939).

von Neumann, John. 1951. »The general and logical theory of automata.« In: *Cerebral Mechanisms in Behavior: the Hixon Symposium*, edited by Lloyd A. Jeffress. 1–41. New York: Wiley.

Wald, Abraham. 1949. »Statistical Decision Functions.« *The Annals of Mathematical Statistics* 20 (2): 165–205.

Weaver, Warren. 1949. »The Mathematics of Communication.« *Scientific American* 181: 11–15.

Wiener, Norbert. 1948. *Cybernetics: Or Control and Communication in the Animal and the Machine*. Paris: Hermann & Cie & Cambridge Massachusetts MIT Press.

Wilkinson, Leland. 2008. »The Future of Statistical Computing.« *Technometrics* 50 (4): 418–435.

Wu, Chien-Fu. 1997. »Statistics = Data Science?«. Accessed July 4, 2023. http://tinyurl.com/barc-studie-datascience.

Zemanek, Heinz. 1971.«Was ist Informatik?« *Elektronische Rechenanlagen* 4: 157–161.

Moral Decision-Making via AI – deep ethics? About shifting or losing responsibility

Tanja Henking

According to the latest statement of the German Ethics Council »Man and Machine – Challenges posed by Artificial Intelligence«, Vice-Chairman Julian Nida-Rümelin stated in a press release that »AI applications cannot replace human intelligence, responsibility and evaluation« (German Ethics Council 2023, press release). In addition to human intelligence, which may still need to be more clearly differentiated from Artificial Intelligence, this refers to two terms in particular, which will be examined closely in the following: evaluation/judgement and responsibility. In essence, it is about evaluation from an ethical perspective, which ultimately also has to stand up to legal scrutiny. This entails responsibility for making the evaluation, the evaluation process and, finally, the outcome of the evaluation/assessment process.

I

The terms »morality« and »ethics« are used in this article, but there is not enough space here to clarify them in greater detail. Put simply, morality is understood as the rules and values generally accepted in a society and ethics as the science of morality. It, thus, also addresses the question of right and wrong. Similarly, the formulation of ethical reflected decisions or actions is used. The question of whether a decision is right or wrong should be separated from the question of what rules and values we give ourselves as a society and which support our coexistence. In this context, the term »evaluation/judgement« mentioned at the beginning presupposes a moral capacity for judgement, which we assume in humans, but which can at best be trained into an Artificial Intelligence (AI). Whether an AI can also learn this is addressed by the question of a moral AI or the concept often already formulated of moral machines (or

only morally acting machines). The question that follows is whether an AI can ever acquire this ability or if it can only ever represent majority ideas. The hint that majority decisions are not always correct, and minorities should not be neglected does not really need any further mention. However, the question of a moral AI ignores a question that needs to be clarified beforehand: the understanding of ethics.

Formulations such as moral or ethical AI fall short. They pretend that the AI »experiences« an ethical conflict and has to solve it. This often involves different questions of morality, ethics and norms, and this on different levels. However, distinguishing between the different levels and tasks is essential if we are to approach the question of a moral AI at all. This begins with the issue of who determines the morals, ethics and norms that are fed into an algorithm and ends with the question of who evaluates the result at the end. Finally, it is also important to know whether humans are able to comprehend the decision-making process at all. This addresses a central (and largely unresolved) question, namely, the explainability of AI (Amann et al. 2020). Does the human consider the result found to be correct or does the AI evaluate it as correct – and what is the benchmark for evaluation? Should we not solve these questions first before we give in to the »temptation« of handing over these difficult questions to AI? And furthermore: Why should we actually want this? What benefits are promised, what risks are feared?

Do humans provoke ethical conflicts by using AI or do they solve them by employing AI? If they want to solve ethical conflicts by means of AI, this is based on an understanding of ethics that should be contradicted at this point. The ethical conflict relates to people and is bound to people. There can be no objective right and wrong, but a person-specific solution is required. The individual involved in an ethical conflict cannot allow themself to be relieved of responsibility by an AI (Sparrow 2021). This will be shown in the following.

This is because for an AI to »experience« a conflict, it must not only be able to decide between A and B, but also take into account the values that shape the situation and influence a decision, it must see what is in favour of A and in favour of B. In the case of a real ethical conflict or even dilemma, we see humans struggling with it, wrestling to make the right decision, and perhaps even suffering from the fact that there will always be vulnerabilities left in the decision-making process under any consideration. When humans have gone through this conflict, wrestled with a decision, we will often be able to accept

the decision made. An AI will not do this; it makes the decision according to a predetermined principle[1] (Iphofen and Kritokos 2019).

II

This raises the question of whether AI has morals or can be trained to have them. What moral concepts underlie the AI? What ethical values does the AI pursue? This question is anything but trivial. If an AI is to be used for questions that are primarily normative in nature, the search for an answer to this question is almost mandatory, because only then it is possible to make transparent according to which ideas the AI is operating, and which value system has been »written« into its training data set. If one takes a norm from a law, then the AI would at least have to be able to consider the legal methods of interpretation, and feed in case law and the literature of legal science. It would need to be able to classify possible »minority opinions« according to their significance for the debate. So, is the legal machine possible? Why law also needs common sense cannot be part of this contribution, which is to be limited to morality, ethics and responsibility. Nevertheless, some of the following considerations will be transferable.

However, first of all, let us return to moral values. This raises questions that have preoccupied philosophers and ethicists for centuries. According to which principles do we judge our actions? The deontological approach, which is predominant in the German tradition, and a utilitarian approach will be chosen for the following considerations in order to shed light on the problem of the criteria according to which a decision should be made. The ethics of principles, which is widespread in medical ethics, will be used again later to question whether the shifting of decisions to an AI-based system is convincing and which considerations should override the question of mere feasibility and, better still, be placed in front of it.

According to deontology, an action in itself can be good or bad, wrong or right. The action is not evaluated according to its consequences. Following its probably most prominent representative, Immanuel Kant, the action is judged according to whether it is in compliance with an obligatory rule and if the ac-

1 The discussion about the possible legal personality of the AI in the future is not intended to take place here.

tion is committed based on this obligation. Deontology, thus, focuses on the preconditions of action (Stanford Encyclopedia of Philosophy 2020).

By contrast, utilitarianism, which goes back to Jeremy Bentham (1780) and, thus, has its origins in the 18th century, pursues a different idea: action is evaluated according to its utility. Therefore, from a utilitarian perspective, the greatest possible benefit for as many people as possible is the standard for correct human decision-making behaviour (Stanford Encyclopedia of Philosophy 2017). *Prima facie*, it is easier to work with this approach in the context of AI (similar Iphofen & Kritikos 2019). It is important to keep this in mind because we might be abandoning our tradition.

III

Medical decision-making situations, as situations with normative impact and which could affect every human being, will be the focus of the following section. AI-based clinical decision support systems (CDSS) are increasingly finding their way into medicine. The Central Ethics Committee of the German Medical Association (ZEKO) dedicated a statement of its own regarding this issue in 2022, emphasising the role of the human being in addition to the possible benefit, whereby the human being should not be forced out of the decision (ZEKO 2022).

The concern about automated decisions is expressed in Art. 22 of the General Data Protection Regulation, where it states that »the data subject shall have the right not to be subject to a decision which produces legal effects concerning him or her or similarly significantly affects him or her and which is based solely on automated processing […]«.

The CDSS can now help to find the right diagnosis, issue warnings, for example, if there is a risk of impending sepsis, or make therapy suggestions for a previously defined diagnosis.

IV

As part of the Check.App project funded by the Federal Ministry of Education and research (BMBF), the ethical, legal and social implications of so-called symptom checker apps that allow users to classify their symptoms by means of an app-controlled medical history are being investigated (Wetzel et al. 2022;

Müller et al. 2022). The apps work predominantly with the probabilities of a diagnosis. This is followed by the recommendation to consult a doctor or even go to an emergency room. These apps are currently being discussed primarily from the point of view of accuracy (Gilbert et al. 2022). However, it is interesting to see how the use of these apps could affect the doctor-patient encounter in the long term. As long as one sees this app operating in a social insurance system with free health care, a noticeable impact on the health-care system is probably not expected. The fact that the resources of even a fundamentally well-equipped health-care system are finite is shown by the current debate about the congestion of emergency rooms, with the result that work is being done on the use of algorithms for triage, the use of which, however, is not uncontroversial due to fears about patient safety (Elsner et al. 2018; Marburger Bund/DGINA/DIVI, Gemeinsame Pressemitteilung 2021).

The keyword »triage« addresses an area that exemplifies a real ethical dilemma. First of all, triage means sifting, sorting. The term originally comes from disaster medicine. The objective of triage is to save as many people as possible with limited resources. Whereas before the COVID-19 pandemic the term was probably only known to experts, the pandemic has turned it into a description of a scenario in the general population which people faced with worry and fear. While German hospitals – to the best of our knowledge – do not have to triage in the same way as hospitals in France, New York or Bergamo/Italy, debates have, nevertheless, been triggered that have, at least, called into question certain basic ethical assumptions (German Ethics Council 2020). The discussion was initially reduced to the question of who should receive a ventilator when only one is available, but two patients need it. The question is broadened to include the situation of the patient who arrives later and for whom no ventilator is available, but who may have a better survival situation than the patient who has already been ventilated. If we take the situation that is initially easy (or easier) to assess from a legal point of view: more patients arrive at the same time and there is not enough capacity available. Not all patients can be treated and saved (or given a chance of being saved). If the doctor saves patient A but not B, the doctor cannot be blamed from a legal point of view. This is because two obligations meet here, both of which the doctor cannot fulfil. Which patient he or she chooses, A or B, is ultimately irrelevant. In order to relieve physicians of the burden of this difficult and stressful decision-making situation, the medical society of intensive care physicians has provided them with criteria to help them make decisions. These criteria focus primarily on the prospects of a successful treatment (DIVI

2020). In other words, those who have a better chance of surviving the disease through treatment should receive it. Those who are less likely to survive, such as those who are older, overweight or have certain pre-existing conditions, will not receive the treatment. The idea of transferring this decision-making algorithm to an AI is not far-fetched, and it has sometimes been reported that such an algorithm has been used. If one considers the idea of triage according to the criteria of success, it quickly becomes apparent that the distribution of limited resources is based on the greatest possible benefit. Triage is utilitarian by its very nature. The medical likelihood of success may be a selection criterion, but it is not mandatory. Therefore, other criteria have been put up for discussion in the context of the debate on triage in the pandemic. These range from age and, thus, life expectancy (Hoven 2020), the juxtaposition of the young mother versus the older man with a previous illness, to the decision by lot (Walter 2022). If one were to (be able to) agree on one of these criteria, an AI would also have to take this into account if it is to decide morally in the sense of a normative consensus. However, this type of evaluative criteria, which can also be classified as utilitarian, has so far been consciously avoided in the German legal system. The decision of the Federal Constitutional Court on the Aviation Security Act (Luftsicherheitsgesetz), which was passed with the collapsing Twin Towers in New York in mind, is groundbreaking. Can shooting down a plane carrying 300 people be justified if it could save 3000 people? The Federal Constitutional Court has rejected the quantification of human lives and declared the authorisation to shoot down an aircraft unconstitutional and not justified (BVerfGE 115, 118 – 166). From a legal point of view, one thing is important to distinguish: the Aviation Security Act would have provided an authorisation basis for the selection decision to the detriment of the aircraft occupants. If the pilot themself were to make a decision in the acute, specific situation that – regardless of how they decide – leads to the death of human lives, then they would not be blamed for making a particular decision. Similar to the case of medical triage, if all the people cannot be saved, the doctor cannot meet all the obligations, so cannot be blamed for their selection decision (Gerson 2020). That their decision remains a personally burdensome one, because their rescue decision is, at the same time, a non-salvation decision, is the essence of the real ethical dilemma. The person remains (morally) responsible for their decision.

Similar scenarios have recently been discussed for autonomous driving. Which group of people should the car steer into if a collision is unavoidable: the group of children or the group of pensioners? Here too, similar to the real triage

described above, a genuine ethical dilemma is constructed, which is, however, provoked by the use of autonomous vehicles. Therefore, it is demanded in some cases that the human being must remain (in the loop). Human beings can act neither rightly nor wrongly, because no matter how they decide, their decision will not be entirely without harm to one person or one group of people and, simultaneously, of benefit to the other person or group of people. Would chance, therefore, be the only ethically correct decision or should another ethically correct decision be taken into account within the framework of a programmed decision and would this decision then be dependent on the respective value system of a society. So should the car have been driven into group A in one country and into group B in another? This would be a normative question that would have to be decided in advance of an anticipated collision. However, this shows one thing quite clearly here: whereas the situation in the triage described above falls on the doctor and they have to make a decision, this triage is one that is calculated in advance. The algorithm will experience a moral stress and no decision burden. It is not a decision that affects or concerns the algorithm personally. It suffers from neither the decision to be made nor the consequences. For the algorithm, the decision remains impersonal. This makes one thing clear: ethical conflict is personal (Sparrow 2021). It is, therefore, not a question of an ethical AI or ethical decision-making in these situations, but rather of an anticipated, ethically problematic decision-making situation for which a decision has already been made in advance of the actual situation.

However, this raises the question of whether the use of algorithms/AI would reduce or even increase the decision-making burden and the moral distress associated with it. Can the person hand over the decision and, thus, relieve themself, or would they experience it as a burden and ultimately feel powerless? Another difficulty in this context, however, should be highlighted. The starting point was the moral or ethical question underlying a decision. Although the law represents a basic moral consensus of a society, it can also raise ethical questions. When the legislator passes a norm, however, it is based on fundamental decisions of coexistence, on the one hand, and requires interpretation and is subject to further development, on the other hand, as advanced by the literature in the legal sciences and case law. Concepts in need of interpretation such as »well-being« come to mind. Well-being can be understood subjectively or objectively (Braun et al. 2022). While the »best interests« principle applies in other legal systems, German guardianship law, for example, does not recognise this, and it is not uncommon for translational inaccuracies and difficulties of interpretation to arise here in legal comparisons (relevant,

inter alia, in decisions concerning the end of life). Moreover, even the question of what is in a person's best interests is subject to interpretation and will be influenced by changing social understanding and Zeitgeist. Incidentally, the legislator decided to delete the term »well-being« in the law on guardianship in order to counter difficulties of interpretation and misunderstandings observed in practice (Schnellenbach et al. 2020; Henking, 2022). The legislator has now also enacted a law in the area of triage within the framework of the Protection against Infection Act. To the best of our knowledge, this regulation on triage is unique throughout the world. The legislature was called upon to act after the Federal Constitutional Court (BVerfG, decision of 16.12.2021, ref.: 1 BvR 1541/20) shared the concerns expressed in a constitutional complaint of discrimination against persons with disabilities in the case of triage. The German Protection against Infection Act was expanded by Sec. 5c of the act to include a prohibition of discrimination, according to which no one would be disadvantaged in the event of the insufficient availability of intensive care treatment capacities essential for survival because of, *inter alia*, their disability, age or degree of frailty. The allocation decision should only be made based on the current and short-term survival probability of the patient concerned. This already raises the question of how short »short-term« should be defined. Comorbidities may be taken into account, but only in the assessment of the current and short-term probability of survival if they significantly reduce the short-term probability of survival related to the current disease due to their severity or combination. If one were to make this decision (which, according to the law, has to be made by two doctors) with an AI-based decision support system, then all these normative preconditions, which are controversial in detail, would have to be taken into account. It would be feasible to limit the AI to a statement on the probability of survival, which only hints at another difficulty: the uncertainties of any prognosis – especially in the case of people, all of whom have a poor prognosis to begin with. Just consider that less than half of the patients suffering from COVID-19 who received invasive ventilation have survived (Karagiannidis et al. 2020). One can imagine AI-based decision support systems in various clinical settings where they are already partially in use (see above). But if one imagines this system in end-of-life decisions, then the question arises here, too, whether ethical, normative questions are excluded or frozen to a value concept at time X. And once again, the question arises as to which moral and value concepts influence the decision. This is because, according to our system of values, we do not follow a »best interest« principle (the content of which would also have to be clarified) which would

allow for an objective consideration, but what is required is a determination of the (presumed) will of the individual oriented towards their subjective interests. In other words, the value system of the person, of the individual, has to be used for the decision of further treatment or the limitation of therapy. This, in turn, makes the ethical decision a personal decision, both for the individual concerned and the person who may have to make it on their behalf. Let us assume that this decision has to be made by a relative (health-care proxy), then this decision remains a personal decision (Sparrow 2021). The relative has to determine and consider the values of the person concerned in their own individual case. They cannot remove themselves from this responsibility; they cannot transfer it to an AI system (Sparrow 2021). The AI will not be able to tell them what their relative would have wanted in concrete terms or whether they themselves are struggling with the interpretation and implementation of this will, because their decision also has consequences for their own life. Other considerations could possibly run the risk of giving (too much) weight to general values. The ethical complexity here arises from the challenge of determining the will of the person in the concrete situation and, thus, making the »right« decision for them.

The temptation to use AI for such decision-making situations may seem attractive. Meier et al. (2022) have presented a proof of concept for ethical decision-making in the clinic. This is intended to help with ethical decisions. Their main aim is to show whether a decision support for ethically difficult questions can be developed by means of an algorithm. The authors obtained their training data set from clinical cases of clinical ethics committees. The cases are categorised in the areas of abortion, decisions regarding minors, refusal of treatment and end-of-life decisions, among others. They used the principles of ethics according to Beauchamp and Childress (2019), the approach most frequently used in the clinical field to discuss ethical questions. They have tried to operationalise these principles of ethics and develop a corresponding algorithm. Their approach showed an astonishingly frequent agreement between the algorithm-based recommendations and those of experts or textbook cases. But is this now also proof that ethical questions can be solved in an AI-based way? And why would one want to do this? The authors speak, among other things, of limited time and personnel resources, which, in the worst-case scenario, could again mean triage, even if these case constellations did not play a role in the data set of the research design. However, the question remains: what are we actually hoping for (see critical statement by Gundersen and Baroe 2022)? What is the understanding of ethics behind these considerations (Spar-

row 2021)? Does the use of AI lead to a reduction in the workload of the staff responsible for the decision-making and care of the patient? This can only be convincing if the relinquishment of decision-making authority reduces stress. Can the doctor be relieved of responsibility? Can it shift the decision? However, the person remains the one who implements the decision in the end. They remain involved and, thus, it is still a personal ethical decision. The idea of ethical case consultation also aims to show the different values, the various interpretations and their points of contact that exist in a group and to work out these different viewpoints in a moderated process and develop a willingness to take on the perspective. This deliberation cannot be done by an AI. It remains a human task. The decisions that have to be made in the context of an ethical case conference represent ethical problems. Rarely do we encounter real ethical dilemmas as in triage. The authors are rightly confronted with the criticism that they have not sufficiently explained why the use of AI is necessary (Sauerbrei et al. 2022). Both the WHO warning not to use AI as a first resource and the EU Ethical Guideline for Trustworthy AI (HLEG 2019) have to be recalled in this context.

If we once again combine the examples of autonomous driving and decision-making in medical contexts, the question remains open as to whether and when we are prepared to leave or hand over the decision to an AI. At what level of automation does the person withdraw from the responsibility for the decision or is there a pressure to justify wanting to decide differently from the AI (the comparison of autonomous driving can be found at Eric Topol 2019; Henking, 2023; Duttge 2019; Katzenmeier 2019)?

V

While AI can be expected to bring new accuracy, an increase in knowledge and objectivity to the assessment of medical prospects of a successful outcome, in addition to the finding that the human body reacts individually, it is first necessary to address where and when implicit values play a role. Decisions at the end of life are a good example of this, especially because the increasing need as a person to be able to make decisions as independently as possible at the end of life has played an increasingly important role in recent decades. This also reflects the values of a society whose morals and ethics are changing. Whether this individuality and development can be depicted sufficiently quickly by an AI must, at least, be doubted. This is because it always requires sufficient reflec-

tion and a discourse on ethical ideas. Society cannot pass on the responsibility for this discourse to an AI, but must take it on itself as an ongoing task. This addresses part of the question of what constitutes ethics. Ethical conflicts can arise from the use of an algorithm, as in the inevitable collision in autonomous driving, because a situation is anticipated and the outcome is already determined. These considerations can be transferred to CDSS depending on how much the system displaces humans from the decision-making process. This suppression must find its limits when individual value concepts, and thus ethical questions, affect the person in terms of their own value system and ethical ideas. The ethical conflict is a personal one that can neither be left to an AI nor for which a proxy can absolve itself of responsibility (Henking, 2023; Sparrow 2019). He or she may even experience a hypothetical AI decision as powerlessness vis-à-vis the system.

VI

Before speaking of moral, ethical algorithms (in clinical use), we should reflect on what kind of ethics we are talking about. Is it the use of algorithms/AI where ethical problems can be anticipated? Should algorithms/AI be used to solve ethical problems? Should the algorithm itself act morally? Especially when answering the last question, it is necessary to think more deeply about what moral decision-making means and what competences, including the struggle for right and wrong, this requires. The evaluation of right and wrong cannot be left to an AI – it is the responsibility of humans and should remain so in the future. Those who call an AI ethical should always declare their understanding of ethics and morality.

References

Amann, Julia, Alessandro Blasimme, Effy Vayena, Dietmar Frey, and Vince I. Madai. 2020. »Explainability for artificial intelligence in healthcare: a multidisciplinary perspective« *BMC Medical Informatics and Decision Making* (20): 310, DOI: 10.1186/s12911-020-01332.

Alexander, Larry, and Michael Moore. 2021. »Deontological Ethics« *The Stanford Encyclopedia of Philosophy* (Winter 2021 Edition), edited by Edward N. Zalta.

Accessed April 22, 2023. https://plato.stanford.edu/archives/win2021/entries/ethics-deontological/.

Bentham, Jeremy. 1870. *An introduction to the principles of morals and legislation*. (1907 reprint of 1823 Edn.), Oxford: Claredon Press.

Beauchamp, Tom L., and James F. Childress. 2019. *Principles of biomedical ethics*. (8th ed.), New York: Oxford University Press.

Braun, Esther, Jakov Gather, Tanja Henking, Jochen Vollmann, and Matthé Scholten. 2022. »Das Verständnis von Wohl im Betreuungsrecht – eine Analyse anlässlich der Streichung des Wohlbegriffs aus dem reformierten Gesetz« *Ethik in der Medizin* 2022 (34): 515–528.

Deutscher Ethikrat. 2023. *Mensch und Maschine – Herausforderungen durch Künstliche Intelligenz*. Stellungnahme, Berlin. Accessed April 22, 2023. https://www.ethikrat.org/fileadmin/Publikationen/Stellungnahmen/deutsch/stellungnahme-mensch-und-maschine.pdf.

Deutscher Ethikrat. 2020. *Solidarität und Verantwortung in der Corona-Krise*. Ad-hoc Stellungnahme, Berlin. Accessed April 22, 2023. https://www.ethikrat.org/fileadmin/Publikationen/Ad-hoc-Empfehlungen/deutsch/ad-hoc-empfehlung-corona-krise.pdf.

DIVI. 2021. *Entscheidungen über die Zuteilung intensivmedizinischer Ressourcen im Kontext der COVID-19-Pandemie*. Version 3. Accessed April 22, 2023. https://www.divi.de/joomlatools-files/docman-files/publikationen/covid-19-dokumente/211125-divi-covid-19-ethik-empfehlung-version-3-vorabfassung.pdf.

Driver, Julia. 2021. »The History of Utilitarianism«, *The Stanford Encyclopedia of Philosophy* (Winter 2022 Edition), edited by Edward N. Zalta and Uri Nodelman. Accessed April 22, 2023. https://plato.stanford.edu/archives/win2022/entries/utilitarianism-history/.

Duttge, Gunnar. 2019. »›Decision-Support-System‹ für Therapieentscheidungen am Lebensende?« *Medizinrecht* 37: 771–776.

Elsner, Christian, Martin Blaschka, and Martin Kleehaus. 2018. »App-basierte Systeme im Bereich der medizinischen Notfallversorgung.« *Notfallmedizin up2date* 2018 13(03): 251–266.

EU HIGH-LEVEL EXPERT GROUP ON ARTIFICIAL INTELLIGENCE. 2019. *Ethics Guideline for Trustworthy*. Accessed April 22, 2023. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai.

Gerson, Oliver Harry. 2020. »§ 3 Pflichtenkollision«. In *Pandemiestrafrecht*, edited by Esser, Robert and Michael Tsambikakis, C.H.Beck, München.

Gilbert, Stephen, Alicia Mehl, Adel Baluch et al. 2020. »How accurate are digital symptom assessment apps for suggesting conditions and urgency advice? A clinical vignettes comparison to GPs.« *BMJ open* 2020 10 (12): e040269.

Gundersen, Torbjørn, and Kristine Bærøe. 2022. »Ethical Algorithmic Advice: Some Reasons to Pause and Think Twice« *The American Journal of Bioethics* 22 (7): 26–28.

Henking, Tanja. 2023. »Theorie- und evidenzbasierte Gesundheitspolitik in Zeiten der Digitalisierung und Künstlicher Intelligenz. Ethische und rechtliche Überlegungen.« In: *Smart Regulation: Theorie- und evidenzbasierte Politik*, edited by Wendland, Matthias, Iris Eisenberger, and Rainer Niemann, 1–13.

Henking, Tanja. 2022. »Die Reform des Betreuungsrecht« In: *Der Nervenarzt* 2022 93: 1125–1133, DOI: 10.1007/s00115-022-01355-6.

Hoven, Elisabeth. 2020. »Die »Triage«-Situation als Herausforderung für die Strafrechtswissenschaft«. In: *Juristenzeitung* 2020: 449.

Iphofen, Ron and Mihalis Kritikos. 2019. »Regulating artificial intelligence and robotics: ethics by design in a digital society« *Contemporary Social Science* 16 (2): 170–184 DOI: doi.org/10.1080/21582041.2018.1563803.

Karagiannidis, Christian, Carina Mostert, Corinna Hentschker et al. 2020. »Case characteristics, resource use, and outcomes of 10 021 patients with COVID-19 admitted to 920 German hospitals: an observational study« *Lancet Respir Med* September 2020 8(9): 853–862, DOI: 10.1016/S2213-2600(20)30316-7.

Katzenmeier, Christian. 2019. »Big Data, E-Health, M-Health, KI und Robotik in der Medizin. Digitalisierung des Gesundheitswesens – Herausforderung des Rechts« *Medizinrecht* 37: 259–271.

Koch, Roland, Anna-Jasmin Wetzel, Christine Preiser et al. 2022. »Ethical, legal, and social implications of symptom checker apps in primary health care (Check.App): Protocol for an interdisciplinary mixed methods study« *JMIR Research Protocols* 11 (5): e34026.

Marburger Bund, DGINA, DIVI.2021. *Gemeinsame Pressemitteilung, Keine Experimente mit der Patientensicherheit!* Accessed April 22, 2023. https://www.marburger-bund.de/sites/default/files/files/2021-02/Gemeinsame%20Pressemitteilung%20MB%2C%20DGINA%2C%20DIVI%20-%20Keine%20Experimente%20mit%20der%20Patientensicherheit.pdf

Meyer, Lukas J., Alice Hein, Klaus Diepold, and Alena Buyx. 2022. »Algorithms for Ethical Decision-Making in the clinic: A proof of concept« *The American Journal of Bioethics* 22 (7): 4–20.

Müller, Regina, Malte Klemmt, Hans-Jörg Ehni et al. 2022. »Ethical, legal, and social aspects of symptom checker applications: a scoping review« *Medicine, Health Care and Philosophy* 25 (4): 737–755.

Sauerbrei, Aurelia, Nina Hallowell, and Angeliki Kerasidou. 2022. »Algorithmic Ethics: A Technically Sweet Solution to a Non-Problem« *The American Journal of Bioethics* 22 (7): 28–30.

Sparrow, Robert. 2021. »Why machines cannot be moral« *AI & Society* 36: 685–693.

Topol, Eric. 2019. *Deep Medicine. How Artificial Intelligence Can Make Healthcare Human Again*. New York: Basic Books.

Walter, Tonio. 2022. »Keine Verpflichtung für ein Triage-Gesetz – und kaum Vorgaben dafür«. In: *Neue juristische Wochenschrift* 2022 (6): 363–366.

Zentrale Ethikkommission der Bundesärztekammer (ZEKO). 2021. *Entscheidungsunterstützung ärztlicher Tätigkeit durch Künstliche Intelligenz*. Stellungnahme. Accessed April 22, 2023. https://www.zentrale-ethikkommission.de/fileadmin/user_upload/_old-files/downloads/pdf-Ordner/Zeko/ZEKO_SN_CDSS_Online_final.pdf.

AI-assisted reflection in child welfare

Maximilian Kraus, Jennifer Burghardt, Christopher Koska

Introduction

Recognizing children's rights and the state's responsibility to ensure and promote these rights plays a crucial role in ensuring children's well-being. The most important international human rights instrument for children was created with the UN Convention on the Rights of the Child (United Nations 1989). In countries in which a state responsibility for the protection of children is anchored, different national approaches have been developed to carry out this task. This assignment is often taken on by social workers. Social work is a practice-based profession and an academic discipline which promotes the empowerment and liberation of people based on the principles of social justice and human rights (IFWS 2014). With regard to risk-oriented approaches to social work, the common challenge lies primarily in the diagnosis and prognosis of child welfare hazards (Schrödter 2020). A perfect, faultless decision hardly seems possible. It would require complete information, sufficient time and unlimited cognitive abilities. We are also aware of the distorting influences on human judgement. While the best interests of the child are obviously the guiding principle for professionals, stereotyping, bias and personal experiences of decision-makers can influence their judgment (Gutwald et al. 2021). Considering this, there has been a growing interest in leveraging the potential of Artificial Intelligence (AI) to enhance the decision-making process and promote better outcomes for children and families.

The primary purpose of this article is to explore the role of AI-based reflection in child welfare cases. Reflection in this context refers to the comprehensive analysis and circumstances as well as potential outcomes associated with child welfare cases. By employing AI as a reflective tool, social workers

and decision-makers can augment their judgment, enhance accuracy, and gain deeper insights into the dynamics of each case.

The article aims to shed light on the potential benefits and challenges of integrating AI into child welfare practices, with a focus on how AI-based reflection can positively impact the field. It will examine AI applications in data collection, analysis, and decision-making, while emphasizing the importance of transparency and accountability in this domain. Additionally, statements in this article will be underpinned by the findings of the KAIMo project.

The article will argue that it is essential to recognize that while AI has the potential to offer valuable support in child welfare cases, it is not a substitute for human involvement and judgment. Therefore, the scope of this article includes an exploration of how AI can complement and enhance existing child welfare practices while recognizing the unique qualities that social workers bring to their roles. The vision is a harmonious interplay between human expertise and AI-generated insights, which together can provide a more robust, comprehensive framework for safeguarding children and families.

Furthermore, this article goes into specific ethical concerns related to the deployment of AI in sensitive contexts such as child welfare. It emphasizes the multifaceted nature of bias in AI systems, highlighting the necessity to examine and address biases at various levels – from information acquisition and data collection to the interpretation of results. AI not only has the potential to enhance efficiency and accuracy but also enables the exposure of existing prejudices.

The limitations of AI-based reflection in child welfare will also be discussed, including challenges related to data availability, data security, potential resistance to technological adoption, and the need for continuous monitoring and improvements of AI algorithms.

In summary, this article aims to provide an application-oriented guideline to AI-based reflection in child welfare cases. By highlighting the potential benefits, ethical considerations, and challenges, this research seeks to foster informed discussions and responsible implementations of AI technologies in the child welfare domain. As AI continues to advance, understanding its role in supporting child welfare endeavors becomes increasingly crucial in the pursuit of a safer, more nurturing environment for children and families in need.

Understanding Child Welfare Cases

Definition of Child Welfare in Germany

In Germany, the definition of child endangerment, as specified in the case law of the Federal Court of Justice, forms the basis of state action in child protection. Accordingly, the well-being of the child is endangered »if a current danger is identified that is of such magnitude that, as things develop, there is a reasonable likelihood of significant damage to the child's mental or physical well-being« (BGH 2016).

However, determining that a child is endangered does not automatically legitimize state measures to protect children. In Germany, the care and upbringing of children is primarily the natural right of parents and their primary duty (Art 6 (2) GG). It is therefore initially the task of the parents to maintain or restore the protection of the child. The basis for state action and thus intervention in the rights of parents is only given if a child's well-being is found to be endangered and the parents are unwilling or unable to avert these dangers themselves.

Importance of Accurate Reflection

Reflective practice is a fundamental aspect of effective social work and child welfare. It involves the critical examination of one's actions, decisions, and assumptions in the context of the cases being handled. In particular, the requirement for forecasting future developments and the long-term effects of decisions represents a major challenge in the decision-making process of professionals. This is further intensified by aspects such as time pressure and pressure to act, lack of (human) resources, lack of information and the personal responsibility of professionals for the consequences of the decision. These conditions further complicate the already complex task of decision-making (Kinderschutz-Zentrum Berlin, 2009).

Therefore, in addition to a precise diagnosis and risk assessment, collegial case counselling and self-reflection are also required before a decision is made. While collegial case counselling serves to systematically analyze the situation and develop options for action in a goal-oriented manner, self-reflection focuses on analyzing one's own attitude and emotions, e.g. towards the parents and the child. This allows professionals to gain a deeper insight into the com-

plexity of the situation in question and recognize possible personal biases or blind spots that may influence their judgement.

Procedures in Risk Assessments

Various methods and instruments are available to support the decision-making process of the professionals. The risk assessment tools help with the systematic collection and assessment of relevant information. They differ primarily in how the information is evaluated (Schrödter 2020). The different instruments can be classified into two basic categories. On the one hand there are statistical forecasting methods that are based on empirical knowledge of certain characteristics and behaviors that apply specifically to the target group of children and young people at risk. To ensure easy manageability, these are mostly used in the sense of checklists with fixed weightings of individual factors. On the other hand, there are interpretative, hermeneutic forecasting methods in which the domain knowledge of the professionals plays a major role. The evaluation of information usually takes place in the discourse of several actors (Lehmann 2024).

Regardless of the debate about AI, the statistical forecasting methods in particular are critically discussed in the technical discourse. Critics of these instruments argue, among other things, that (group) statistical predictions cannot be applied to individual cases and that categorizing facts does not serve the actual »understanding of the case«. »In this sense, professional action is not characterized by technology orientation and dogmatic adherence to rules, but by an understanding of the case for which scientific knowledge is only one necessary element. This must be supplemented by empirical knowledge and hermeneutic sensitivity to the case« (Ferchhoff and Kurz 1998, p. 23).

The Federal Association of Child Protection Centers e.V. (Bundesarbeitsgemeinschaft der Kinderschutz-Zentren e.V.) also criticizes the use of risk assessment forms. »The use of the questionnaires represents a quality risk in the risk assessment and encourages professional errors«, for example if the design of the questionnaires »a) suggests objectivity and certainty, b) leads to a binding result on its own (points system, traffic light system) and thus leads to the questionnaire and no longer the person making the decision (in the extreme in the sense of a PC software-based evaluation)« (Die Kinderschutz-Zentren 2011, p. 3).

Advocates of statistical prediction methods, on the other hand, emphasize the transparent and verifiable weighting and the influence of certain charac-

teristics on the result of the risk predictions and refer to the numerous individual and meta-studies that prove a significantly higher prediction quality of statistical methods compared to, for example, hermeneutic or clinical methods (Johnson et al. 2015, van der Put et al. 2017, Baird/Wagner 2000, Søbjerg et al. 2021). Particularly with regard to the assessment of child endangerment, it is argued that »a risk assessment, as envisaged by the legislator in child protection, cannot be carried out as a purely interpretative judgement; it is ultimately a classification process« (p. 9). The use of statistical instruments is therefore »helpful, because they can be used to make accurate predictions and more appropriate group categorizations – both occurances that professionals can only do inadequately on the basis of individual case-related interpretative procedures. Classification is therefore not only not wrong, but first and foremost necessary and common. It is most accurately possible through statistical procedures.« (p. 19ff).Given the ongoing development of AI, it is foreseeable that statistical methods will become established (Gutwald et al. 2021, Burghardt and Lehmann 2023). »It will probably be undisputed in the future that computer-aided forecasts are more accurate than forecasts generated by specialists without the support of computers. In the future, it will no longer be a question of the effectiveness of statistical forecasting methods, nor will it be a question of whether social work using statistical methods should allow the practice to take hold at all.« (Schrödter 2020, p. 255ff). In the following chapters, we will critically examine the integration of AI in child welfare risk assessments, highlighting its potential to enhance decision-making while acknowledging the vital role of human expertise.

Artificial Intelligence in Social Services

Overview

Artificial Intelligence refers to the simulation of specific aspects of human intelligence in machines that are programmed to process and analyze data and make decisions based on that analysis. Though these machines ›learn‹ from patterns and experiences in data, it is important to distinguish that they do so without the conscious or subjective feelings humans tend to have when learning from experience. With advancements in machine learning, natural language processing, and data analytics, AI has gained significant traction in various domains, including healthcare, finance, and customer service, where it

has demonstrated the ability to streamline processes, improve decision-making, and optimize resource allocation. In recent years, the potential of AI to enhance social services has also been explored more thoroughly, with a growing interest in leveraging AI technologies to address the complexities of child welfare cases as well as predictive policing in law enforcement (Mugari & Obioha 2021).

In the context of social services, AI can be used to support decision-making of social workers and other professionals involved in child welfare. AI technologies, such as machine learning algorithms and natural language processing, enable systems to analyze vast amounts of data quickly, detect patterns, and generate insights that can inform decision-making of social workers.

Applications

AI can potentially be applied at various stages of the child welfare process to support social workers and enhance the overall efficacy of interventions. Some of the key applications of AI in child welfare include:

Data Collection and Analysis

AI can streamline the data collection process by automating the extraction of relevant information from various sources, including case files, reports, and historical data. Through advanced data analytics, AI can identify patterns and trends that may not be apparent to human analysts, thus facilitating a more comprehensive understanding of each case (Waldfogel 2000).

Risk Assessment

AI-powered risk assessment tools can assist social workers in evaluating the level of risk to a child's safety and well-being within their family environment. By analyzing a broad range of risk factors and protective factors, these tools can help anticipate future developments, prioritize cases based on urgency and allocate resources more efficiently (Gillingham 2006).

Decision-making Support

AI can serve as a decision-making support system, offering factual insights and recommendations to social workers when considering different intervention strategies. This collaborative approach can augment decision-making practices based on standardized templates and enables social workers to

make more informed and data-driven decisions while maintaining their professional expertise and judgment (Heggdalsvik et al. 2018).

Data collection and Analysis

In child protection, it is crucial for professionals to get a direct impression of the child and its personal environment. All information relevant to the risk assessment should be collected as comprehensively as possible. Based on knowledge of specific risk and protective factors for the child's well-being, systematic observations are carried out and statements are obtained from various actors (e.g. family members, educators/teachers) (Bathke et al. 2019). It must be considered that both personal observations and third-party statements are not only influenced by the subjectivity of the observers but are also shaped through interpretative skills, enabling the integration of collected experiences into a meaningful context, thus go beyond mere subjectivity.

Understanding that data in child protection is an abstraction and can never fully capture the complexity of reality, it is essential that professionals are equipped not just with access to relevant information but also with the methodological and technical means for its appropriate analysis. This section highlights some technical and methodological possibilities for utilizing data within the confines of legal stipulations, while also incorporating preliminary ethical considerations in data handling.

Following the exploration of the technical and methodological possibilities (›What can I do with data?‹ as per the Data Literacy Charter, Schüller et al. 2021), AI technologies offer opportunities to streamline data collection processes and enhance the quality and comprehensiveness of information gathered. A key method of potential AI-based data collection in child welfare is Natural Language Processing (NLP):

NLP algorithms enable systems to analyze unstructured data, such as interview transcripts and case narratives, extracting essential details and identifying patterns in language that may indicate potential risks or concerns (Chowdhary 2020). The technology is capable of processing large data sets, providing relevant information to social workers in a time-efficient manner.

Ethical Considerations in Data Handling

AI-based data collection leads to a complex net of ethical considerations, which are closely linked to legal issues and are intertwined in current regulatory efforts. As a result, many people find it difficult to distinguish genuine

ethical considerations from legal or technical issues (Filipović and Koska 2019). In practice, this often leads to ethical considerations being reduced solely to data protection aspects (such as informed consent, purpose limitation, data minimization, transparency etc.) or to specific measures for organizational and technical assurance of information security (especially institutional responsibilities and competencies).

The EU AI Act stands, after the EU GDPR, as another testament to the EU's commitment to highlighting this linkage. As important and valuable as these efforts are, this approach risks that ethical deliberation only takes place in selected expert circles.

Considering the risk classification of the AI Act, the topic of AI assistance to support reflection in child welfare cases, as discussed by Colloseus in this volume, is to be classified as a high-risk application with regulatory requirements. Such a classification not only demands human oversight of the IT system by an expert who can press a stop button at any time, but also the involvement of all affected parties (Koska and Filipović). However, the ethical dimension of the requirement to involve those affected in the data collection and usage process goes far beyond the legal cornerstone of informed consent. From a philosophical perspective, the central challenge is to implement the aspect of ›meaningful human control‹ not only for a High-Level Expert Group (HLEG) – as proposed in the AI Act – but also for social work professionals as well as the affected families and especially the affected children. After all, no one has such privileged access to their own wishes, goals and interests as the people concerned (Koska 2023).

Analyzing and Extracting Insights from Data

AI-powered data analysis has the potential to transform the field of child welfare by enabling social workers to derive valuable insights from vast amounts of information quickly and efficiently. Several key techniques are employed for data analysis in this domain.

One crucial technique is Pattern Recognition, where AI algorithms can identify patterns and correlations within the data that might not be immediately evident to human analysts. More advanced LLMs (Large Language Models) can analyze file cases, and automatically find hints for or against risk status of a child. Most crucially, they have shown emergent abilities in reasoning about whether certain hints contradict or support each other (Wei et al. 2022). This aids in identifying risk factors and protective factors associ-

ated with child welfare cases, allowing social workers to make more informed decisions.

Although prone to bias, Sentiment Analysis is another valuable technique, using Natural Language Processing (NLP) to assess the sentiments and emotions expressed in written or verbal communications (Mao et al. 2023). This helps social workers better understand the emotions and perspectives of the children and families involved, enhancing their ability to offer appropriate support.

Enhancing decision making

AI systems driven by data are being utilized more frequently to enhance human decision-making in intricate social situations, such as those found in social work or the legal profession (Kawakami et al. 2022b). AI can support social workers to make informed decisions and tailor interventions to suit the unique circumstances of each case, ultimately leading to better outcomes for vulnerable children (Lysaght et al. 2019).

The dimensions of AI-assisted decision-making in child welfare encompass various critical aspects:

AI-driven risk assessment tools have the potential to analyze a wide array of risk and protective factors associated with a child's situation, providing social workers with a comprehensive evaluation of the level of risk to their safety and well-being (Kawakami et al. 2022a). For that, guidelines like the framework for algorithmic decision-making adapted for the public sector (ADMAPS) can help in the design of high-stakes algorithmic decision-making tools in the child-welfare system (Saxena et al. 2021).

Additionally, studies in the field of healthcare have shown that AI's predictive modeling can anticipate potential outcomes that may arise from a given set of risk factors (Axelrod and Vogel 2003). Mapping these approaches to risk factors in a child's life can help enabling social workers to implement timely and preventive interventions. Early identification of concerns can significantly reduce the likelihood of more severe issues developing later on, offering a proactive approach to child welfare.

In conclusion, AI-generated insights can complement social workers' expertise by providing data-driven information that informs their decisions. This collaborative approach empowers social workers to make well-informed choices while preserving their professional judgment and deep understanding of individual circumstances.

Addressing Bias and Fairness Concerns

Automated decision-making systems (ADM) are often associated with greater objectivity and are therefore – at least in principle – considered inherently fairer than human decision-makers (Hauer 2023). In this context, fairness is typically reduced to the concept of non-discrimination. The primary goal of this approach is to identify biased data sets and avoid distortions in the algorithms used. Various strategies are available for this purpose, such as the use of diverse and representative data, methods for bias detection and correction, Explainable AI, and human oversight.

In our research project on the development of an AI-based assistance system for the assessment of child welfare risks, we found that bias can occur at several points in the assessment, planning and controlling process.

Bias can already occur in individuals who report a suspected child welfare risk to the Youth Welfare Office. These individuals may, for example, keep certain families or groups under disproportionate scrutiny due to prejudice. Another point at which bias can occur is the documentation of case files by social work professionals. These professionals could inadvertently incorporate their own prejudices into the case files, for instance, by highlighting certain information or omitting other. Bias can also play a role in the process of colleague case consultations. Social work professionals may, for example, prefer or reject specific strategies based on prejudice.

Building on the processual view of judgment formation, specific forms of bias can be addressed at particular points in time. In the context of the KAIMo project, nearly 200 potential cognitive biases were screened and assigned to various phases of the decision-making process.[1] With the aim of creating an additional level of reflection for social work professionals, a context-sensitive selection of biases was subsequently displayed to the users for the evaluation of the prototype.

1 The interactive graphic from the Institute for Media and Communication Management at the University of St. Gallen, available online at https://bias-map-vl.web.app, provides a good overview and is useful for an introduction tot he topic.

KAIMo: Overview and Methodology

Project goals and assumptions

The research project KAIMo examines whether and to what extent specific elements of the ethical assessment and decision-making process that guides the actions of youth welfare offices can be encapsulated and translated into algorithms, and whether digital tools may strengthen institutional actions. Particularly due to the complexity of the situation, the legal mandate and the scope of the decisions, the question arises as to whether digital systems can help to make institutional decisions more transparent and well-founded. The research project thus opens a fundamental reflection on the field of tension between social work, ethics and computer science.

The basic assumption of the project is that digital systems can relieve and support the daily work of social workers. Through specific forms of analysis and visual presentation of case data, a faster, more transparent, and more efficient case processing becomes possible.

Initially, it was hypothesized that algorithmic decision support could introduce an objective, data-based dimension to the decision-making processes concerning a child's well-being. However, in the course of the project it has become evident that achieving true objectivity is unfeasible due to the inherent subjectivity of the data involved. Considering this, the project's objective has evolved to utilize algorithmic support primarily as a means to reveal blind spots and resolve inconsistencies among the various contributors within the available data, thereby enriching the decision-making process with more comprehensive and subtle insights.

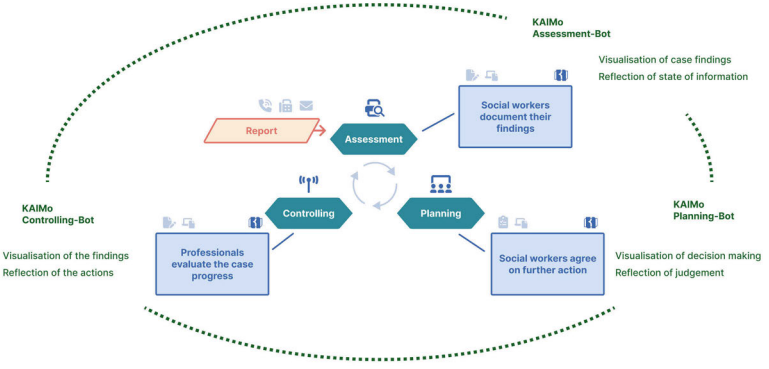
The Three-Agent Approach

Following this guideline, new technologies should not simply be imposed on a social practice, as this can lead to various risks of rejection or improper use.

The KAIMo project worked out an approach to AI integration involving three components (agents) that were designed to ensure a smooth integration in existing processes and established guidelines present in social services. The three agents work together on different levels to provide their respective expertise to the human caseworkers for their case evaluation. They aim to specifically support social work professionals in information gathering (KAIMo Assessment-Bot), decision-making (KAIMo Planning-Bot), and mon-

itoring the implemented measures (KAIMo Controlling-Bot). *Assessment*, *Planning* and *Controlling* are key sub-processes in child welfare services (Figure 1).

Figure 1: Child welfare process



In the first step, existing case data (call logs, notes from professionals, records of home visits) are aggregated and semantically analyzed. The result of this phase is a classified list of case features that provide professional indications for or against the presence of child endangerment.

Building on this classification, the input data is then visualized in a knowledge graph. The structured visual representation of the case characteristics is intended to support professionals in the case analysis and provide case-specific expertise. For example, blind spots can be identified in this phase through a completeness check (comparison with the feature set in similar scenarios), or indications of possible subjective biases of the actors involved can be revealed.

Additionally, context-sensitive reflection questions are asked, derived from the current status and the specific work step of the professionals. These questions are inspired by the idea of the Socratic method, a form of dialogic practice often referred to as ›maieutics‹ or the ›art of midwifery‹, which aims to stimulate critical thinking and illuminate ideas through conversational questioning. For example, the system could ask a question such as »Are there certain assumptions about the family or environment that might be influencing your assessment?« Another question could be »What information is

still missing, and how might these missing pieces affect the overall picture?« to point out possible blind spots in the information situation. In this way, the system supports decision-making by guiding professionals to look critically at their own thought processes rather than providing direct decision recommendations.

Ethical decision-making processes cannot be reduced to merely rational deliberations. Central to a moral consciousness are also abilities such as empathy, intuition, and a form of judgment that presupposes embodiment and the ability to choose for or against the good (Koch 2019; Fuchs 2020; Koska 2023). Since algorithms can at best mechanically replicate such characteristics, it's often misleading to consider them as actors in the context of moral decisions. This is because, given the current state of technology, machines cannot have their own moral values or intentions, as they are incapable of establishing a normative-desiring relationship to a moral decision. Therefore, algorithms should aim to improve human decision-making rather than to replace it and to ensure transparency and comprehensibility of algorithmic decisions. This includes implementing mechanisms to review and adjust algorithms to ensure they adhere to ethical standards and are continuously improved.

Human – AI collaboration

The collaboration of the three agents aims to create a balanced approach to decision-making, taking into account both the efficiency of the algorithms and the ethical integrity of the decisions. However, the human component in this system remains crucial. One central pillar and the main guideline of the KAIMo project is an approach to AI integration that is defined by a supportive role. Predictive analytics models are exclusively used to question and improve the current information status (assessment phase), the process of judgment (planning phase), and the effectiveness of measures (controlling phase). The focus is particularly on making the normative criteria guiding the actions of youth welfare offices visible. In this context, algorithms can serve as assistance to support human decision-makers in ethically complex situations. For instance, they can identify patterns in large data sets that are difficult for humans to grasp, providing valuable information for decision-making. It is important to understand the role of algorithms as supporting and complementing human expertise, rather than replacing it.

Central to the approach followed in the KAIMo project is the belief that both professionals and algorithms bring distinct strengths. When synergized, they

can offer an unparalleled approach to decision-making, tailored to the unique challenges of child welfare. The KAIMo system, with its Knowledge Graph visualization and Socratic model-inspired reflection questions, acts as a supporting arm to social workers, allowing them to make decisions with a broader, data-informed perspective.

Challenges and mitigation strategies

While applications of AI technologies in social services are manifold, as outlined in the previous chapters, it is important to assess benefits and risks of AI adoption in order to evaluate if and how an introduction of this technology can influence everyday work of social workers.

Previous research in other fields such as agriculture and virology has shown its capacity to provide enhanced data insights (Rawson et al. 2019; Gil et al. 2021). By leveraging AI, child welfare agencies could be enabled to conduct more thorough analyses of data, unveiling valuable insights and trends that might be difficult to identify through traditional approaches (Alufaisan et al. 2021).

On the other hand, it's crucial to recognize that AI models can inadvertently inherit biases from the data they are trained on. Addressing bias in computer systems involves three categories, as identified by Friedman and Nissenbaum (1996, pp. 333–336): pre-existing bias rooted in societal structures, technical bias arising from system design, and emergent bias that evolves during real-world use. AI-assisted reflection on bias, as implemented within the KAIMo prototype, can help to reveal pre-existing biases. However, it's important to recognize that the introduction of AI introduces a heightened complexity, particularly on the technical and emergent levels, which may risk displacing or reinforcing bias, making bias mitigation a multifaceted challenge.

Additionally, the use of AI involves the collection and analysis of sensitive information, raising valid privacy concerns. It is imperative to implement data protection measures to safeguard the confidentiality and security of this information. Ensuring that individuals' privacy rights are respected and upheld is challenging and a paramount ethical consideration (Oseni et al., 2021).

Furthermore, the integration of AI in child welfare should not be seen as a replacement for human expertise. Instead, it should be viewed as a tool to augment and enhance the capabilities of human professionals. Collaboration between AI systems and human experts is essential for ensuring the best outcomes for children and families involved in the system (Wang et al., 2020).

In this chapter, we explore the key challenges associated with AI adoption in child welfare as they were recognized as part of the KAIMo project.

Establishing Accountability Mechanisms

Accountability is crucial in child welfare cases as decisions made can have long-lasting effects on the well-being of children and families. Additionally, depending on local law, social workers are oftentimes held accountable as private persons for their decisions and the results thereof (Döring et al. 2023). To establish accountability in AI-driven child welfare systems, it is essential to define clear roles and responsibilities for all stakeholders involved. The KAIMo project implements a multi-agent solution, which involves human social workers as well as AI algorithms.

Additionally, mechanisms for feedback and evaluation should be put in place. Social workers should be encouraged to provide feedback on the AI's performance, and there should be avenues for addressing concerns and correcting errors that may arise during the decision-making process. These accountability mechanisms are vital in ensuring that AI is viewed as a tool to support social workers rather than replace them, maintaining a human-centric approach to child welfare.

Legal and Ethical Implications

The integration of AI in child welfare raises crucial legal and ethical considerations that demand careful attention to safeguard the rights and well-being of children and families involved.

A child-centric approach is imperative, with AI-driven decision-making prioritizing the safety, well-being, and development of children in all recommendations. Nonetheless, human oversight must be maintained, empowering social workers to retain control over AI-assisted decisions and prevent undue reliance on AI recommendations without critical human judgment, as outline by KAIMo's supportive AI approach.

Introducing AI within the realm of child welfare reveals pivotal legal and ethical dimensions that warrant meticulous scrutiny to champion the rights and welfare of both children and their families. At the epicenter of these considerations is the tension between social work and ethics.

Establishing clear lines of accountability and liability is therefore essential, enabling determination of responsibility in case of errors or adverse outcomes resulting from AI-driven decisions. Professionals therefore need the necessary

knowledge of the capabilities and limitations of AI models and the ability to ensure ethical and responsible use of this technology.

One of the foremost concerns in AI-driven child welfare is the security and privacy of sensitive data. To support privacy of sensitive data, AI systems applied in child welfare scenarios should work in a closed system, native to the child welfare institution that applies the AI solution. However, for NLP and machine learning mechanism to work, it is mandatory to grant these processes access to actual case data. After all, the AI should support social workers by giving insightful information and coherences about individual cases.

It was outside the scope of KAIMo to work out detailed guidelines for data handling in child welfare AI applications. These questions deserve dedicated research efforts and should involve legislative and political drivers.

Integration with Existing Systems

Integrating AI systems into the existing child welfare infrastructure can be a complex task. Many agencies already rely on legacy systems, and introducing AI solutions may require substantial changes to the workflow and processes. To address this challenge, the KAIMo project proposes a phased approach to integration.

Initially, AI systems should be introduced in a supplementary role, providing insights and recommendations to social workers without directly replacing any existing processes. This allows for gradual familiarization and adjustment to the new technology. As social workers become more comfortable with the AI's assistance, further integration steps could be taken, streamlining the collaboration between human and AI agents in adherence to the supportive AI approach.

Overcoming Resistance to AI Adoption

Resistance to AI adoption is not uncommon, especially in fields where human decision-making has been the norm for a long time, such as child welfare. End users of AI systems may fear that AI cannot adequately capture and account for the complexity of family systems and life-world influences, disregard professionals' tacit experiential knowledge, or undermine their autonomy in making important decisions. Furthermore, involving social workers in the design and development of the AI system fosters a sense of ownership and trust in the technology. Their feedback is continuously collected and integrated into the AI's improvement processes, ensuring that the system is tailored to meet their specific needs and concerns.

It is also essential to communicate the ethical principles guiding the AI system's development and use to the public. Transparency about how AI is employed in child welfare cases, its limitations, and the steps taken to address potential biases fosters trust and understanding from all stakeholders.

By addressing these challenges and adopting appropriate mitigation strategies, the integration of AI in child welfare cases can be done responsibly and effectively. The KAIMo project serves as a model for how AI can be harnessed to support social workers and improve outcomes for vulnerable children and families while respecting privacy, adhering to ethical guidelines, and maintaining a human-centric approach. As AI technology continues to advance, it is crucial to remain proactive in addressing challenges and ensuring that AI remains a powerful tool in the hands of caring and knowledgeable professionals.

In summary, successfully integrating AI into the child welfare system requires direct actions to ensure its ethical application. This includes employing fairness metrics to evaluate and adjust algorithms, implementing robust bias detection and mitigation strategies, and enforcing stringent data protection measures. Additionally, fostering a collaborative interplay that combines human oversight with AI's analytical capabilities is the most promising approach to enhance decision-making processes. By focusing on these specific and actionable steps, a framework can be created that results in a responsible and beneficial use of AI in this sensitive domain.

References

Alufaisan, Yasmeen, Laura R. Marusich, Jonathan Z. Bakdash, Yan Zhou, and Murat Kantarcioglu. 2021. Does Explainable Artificial Intelligence Improve Human Decision-Making? Proceedings of the AAAI Conference on Artificial Intelligence 35 (8): 6618–6626. doi: 10.1609/aaai.v35i8.16819.

Axelrod, Randy C., and David Vogel. 2003. Predictive Modeling in Health Plans. Disease Management & Health Outcomes 11 (12): 779–787. doi: 10.2165/00115677-200311120-00003.

Baird, C./Wagner, D. (2000): The relative validity of actuarial-and consensus-based risk assessment systems. In: Children and Youth Services Review 22, 839–871.

Bastian, P. (2014). Statistisch Urteilen – professionell Handeln. Überlegungen zu einem (scheinbaren) Widerspruch. Online verfügbar unter: https://www.researchgate.net/publication/265446562_Statistisch_Urteilen_-_professionell_Handeln_Uberlegungen_zu_einem_scheinbaren_Widerspruch

Bundesgerichtshof 2016. Beschluss XII ZB 149/16. Online available: https://juris.bundesgerichtshof.de/cgi-bin/rechtsprechung/document.py?Gericht=bgh&Art=pm&Datum=2016&nr=76862&linked=bes&Blank=1&file=dokument.pdf.

Burghardt, Jennifer, Lehmann, Robert. 2023. Künstliche Intelligenz und Kinderschutz. Dexheimer, Andreas; Rothballer, Marc (Hg.): Künstliche Intelligenz in der Kinder- und Jugendhilfe. Zeitschrift Jugendhilfe, 05/2023, S. 410–414.

Chowdhary, K. R. 2020. Natural Language Processing. In Fundamentals of Artificial Intelligence, ed. K. R. Chowdhary, 603–649. New Delhi: Springer India.

Die Kinderschutz-Zentren (2011). Empfehlung der Kinderschutz-Zentren zur Nutzung von Gefährdungseinschätzungsbögen in den Kinderschutz-Zentren. Verabschiedet vom Fachausschuss der Kinderschutz-Zentren. Köln. Online verfügbar unter: https://jugendhilfeportal.de/fileadmin/user_upload/fkp_quelle/pdf/Empfehlungen%20zur%20Nutzung%20von%20Gefaehrdungseinschaetzungsboegen.pdf

Döring, Linn K., Mörsberger Thomas, Wapler Friederike (eds.). 2023. Garanten für den Kinderschutz? Nomos Verlagsgesellschaft mbH & Co. KG.

Ferchhoff W, Kurtz T (1998) Professionalisierungstendenzen der Sozialen Arbeit in der Moderne, Neue Prax 28(1)12-26.

Filipović, Alexander, Koska, Christopher. 2019. Corporate Digital Responsibility muss mehr als geltendes Recht abbilden: Zu den »Zehn Regeln für den verantwortungsvollen Einsatz von KI in Human Resources« des Ethikbeirats HR-Tech. https://www.future-of-hr.com/2019/10/corporate-digital-responsibility-muss-mehr-als-geltendes-recht-abbilden/.

Friedman, Batya und Helen Nissenbaum. 1996. Bias in computer systems. In: ACM Trans. Inf. Syst. 14 (3), 330–347. DOI: https://doi.org/10.1145/230538.230561.

Fuchs, Thomas. 2020. Verteidigung des Menschen: Grundfragen einer verkörperten Anthropologie.

Gil, Yolanda, Daniel Garijo, Deborah Khider, Craig A. Knoblock, Varun Ratnakar, Maximiliano Osorio, Hernán Vargas, Minh Pham, Jay Pujara, Basel Shbita, Binh Vu, Yao-Yi Chiang, Dan Feldman, Yijun Lin, Hayley Song,

Vipin Kumar, Ankush Khandelwal, Michael Steinbach, Kshitij Tayal, Shaoming Xu, Suzanne A. Pierce, Lissa Pearson, Daniel Hardesty-Lewis, Ewa Deelman, Rafael Ferreira Da Silva, Rajiv Mayani, Armen R. Kemanian, Yuning Shi, Lorne Leonard, Scott Peckham, Maria Stoica, Kelly Cobourn, Zeya Zhang, Christopher Duffy, and Lele Shu. 2021. Artificial Intelligence for Modeling Complex Systems: Taming the Complexity of Expert Models to Improve Decision Making. ACM Transactions on Interactive Intelligent Systems 11 (2): 1–49. doi: 10.1145/3453172.

Gillingham, Philip. 2006. Risk Assessment in Child Protection: Problem Rather than Solution? Australian Social Work 59 (1): 86–98. doi: 10.1080/03124070500449804.

Gutwald, Rebecca, Jennifer Burghardt, Maximilian Kraus, Michael Reder, Robert Lehmann, and Nicholas Müller. 2021. Soziale Konflikte und Digitalisierung Chancen und Risiken digitaler Technologien bei der Einschätzung von Kindeswohlgefährdungen. Ethik Journal 7 (2): 1–20.

Heggdalsvik, Inger Kristin, Per Arne Rød, and Kåre Heggen. 2018. Decision-making in child welfare services: Professional discretion versus standardized templates. Child & Family Social Work 23 (3): 522–529. doi: 10.1111/cfs.12444.

IFWS – International Federation of Social Workers (2014). Global definition of social work. Online available: https://www.ifsw.org/what-is-social-work/global-definition-of-social-work/.

Johnson, W.; Clancy, T. & Bastian, P. (2015): Child Abuse/Neglect Risk Assessment under Field Practice Conditions: Tests of External and Temporal Validity and Comparison with Heart Disease Prediction. Children and Youth Services Review, 56, S. 76–85.

Kawakami, Anna, Venkatesh Sivaraman, Hao-Fei Cheng, Logan Stapleton, Yanghuidi Cheng, Diana Qing, Adam Perer, Zhiwei Steven Wu, Haiyi Zhu, and Kenneth Holstein. 2022a. Improving Human-AI Partnerships in Child Welfare: Understanding Worker Practices, Challenges, and Desires for Algorithmic Decision Support. In CHI Conference on Human Factors in Computing Systems, 1–18. CHI '22: CHI Conference on Human Factors in Computing Systems, New Orleans LA USA. 29 04 2022 05 05 2022. New York, NY, USA: ACM. doi: 10.1145/3491102.3517439.

Kawakami, Anna, Venkatesh Sivaraman, Logan Stapleton, Hao-Fei Cheng, Adam Perer, Zhiwei Steven Wu, Haiyi Zhu, and Kenneth Holstein. 2022b. »Why Do I Care What's Similar?« Probing Challenges in AI-Assisted Child Welfare Decision-Making through Worker-AI Interface Design Concepts.

In Designing Interactive Systems Conference, 454–470. DIS '22: Designing Interactive Systems Conference, Virtual Event Australia. 13 06 2022 17 06 2022. New York, NY, USA: ACM. doi: 10.1145/3532106.3533556.

Kinderschutz-Zentrum Berlin e.V (2009). Kindeswohlgefährdung. Erkennen und Helfen. 11. überarbeitete Auflage. Online available: https://www.bmfsfj.de/resource/blob/94156/178873b3c5a6eeb604568df609e16683/kindeswohlgefaehrdung-erkennen-und-helfen-data.pdf.

Koch, Christof. 2019. The feeling of life itself: Why consciousness is widespread but can't be computed. Cambridge, Massachusetts, London, England: The MIT Press.

Koska, Christopher. 2023. Ethik der Algorithmen. Auf der Suche nach Zahlen und Werten. (1. Auflage 2023.) Berlin: Springer Berlin; J.B. Metzler (Techno:Phil – Aktuelle Herausforderungen der Technikphilosophie, 6). doi: doi.org/10.1007/978-3-662-66795-8.

Koska, Christopher, and Alexander Filipović. 2019. Blackbox AI – State Regulation or Corporate Digital Responsibility? In Digitale Welt 3 (4), 28–31. doi: 10.1007/s42354-019-0208-5.

Lehmann, R. (2024, im Druck). Herausforderungen der künstlichen Intelligenz in der Sozialwirtschaft. In L. Kolhoff (Hg.), *Aktuelle Diskurse in der Sozialwirtschaft V* (S. XX–XX). Springer.

Lysaght, Tamra, Hannah Yeefen Lim, Vicki Xafis, and Kee Yuan Ngiam. 2019. AI-Assisted Decision-making in Healthcare: The Application of an Ethics Framework for Big Data in Health and Research. Asian bioethics review 11 (3): 299–314. doi: 10.1007/s41649-019-00096-0.

Mao, Rui, Qian Liu, Kai He, Wei Li, and Erik Cambria. 2023. The Biases of Pre-Trained Language Models: An Empirical Study on Prompt-Based Sentiment Analysis and Emotion Detection. IEEE Transactions on Affective Computing 14 (3): 1743–1753. doi: 10.1109/TAFFC.2022.3204972.

Mugari, Ishmael, and Emeka Obioha. 2021. Predictive Policing and Crime Control in The United States of America and Europe: Trends in a Decade of Research and the Future of Predictive Policing.

Oseni, Ayodeji, Nour Moustafa, Helge Janicke, Peng Liu, Zahir Tari, and Athanasios Vasilakos. 2021. *Security and Privacy for Artificial Intelligence: Opportunities and Challenges.*

Rawson, Timothy M., Raheelah Ahmad, Christofer Toumazou, Pantelis Georgiou, and Alison H. Holmes. 2019. Artificial intelligence can improve decision-making in infection management. Nature human behaviour 3 (6): 543–545. doi: 10.1038/s41562-019-0583-9.

Saxena, Devansh, Karla Badillo-Urquiola, Pamela J. Wisniewski, and Shion Guha. 2021. A Framework of High-Stakes Algorithmic Decision-Making for the Public Sector Developed through a Case Study of Child-Welfare. Proceedings of the ACM on Human-Computer Interaction 5 (CSCW2): 1–41. doi: 10.1145/3476089.

Søbjerg, L. M., Taylor, B. J., Przeperski, J., Horvat, S., Nouman, H., Harvey, D. (2020). Using risk factor statistics in decision-making: prospects and challenges. European Journal of Social Work, 0(0), 1–14. Routledge.

Schrödter, M., Bastian, P., Taylor B. (2020). Risikodiagnostik und Big Data Analytics in der Sozialen Arbeit. In: Kutscher et al. (Hg.). Handbuch Soziale Arbeit und Digitalisierung. 255–263.

Schüller, Katharina, Henning Koch, and Florian Ramplet. 2021. Data Literacy Charta. Stifterverband (2021): Data Literacy Charta, Online available: https://www.stifterverband.org/data-literacy-charter.

United Nations (20.11.1989). Convention on the Rights of the Child. Online available: https://www.coe.int/en/web/compass/convention-on-the-rights-of-the-child.

van der Put, C. E., Assink, M., Boekhout van Solinge, N. F. (2017). Predicting child maltreatment: A meta-analysis of the predictive validity of risk assessment instruments. Child Abuse & Neglect, 73, 71–88.

Waldfogel, Jane. 2000. »Child welfare research: How adequate are the data?« *Children and Youth Services Review* 705–741.

Wang, Dakuo, Elizabeth Churchill, Pattie Maes, Xiangmin Fan, Ben Shneiderman, Yuanchun Shi, and Qianying Wang. 2020. »From Human-Human Collaboration to Human-AI Collaboration: Designing AI Systems That Can Work Together with People.« *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems* 1–6.

Wei, Jason, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, et al. 2022. »Emergent Abilities of Large Language Models.« *Transactions on Machine Learning (08/2022)* 1–30.

Addressing the needs and demands of child welfare: A connection between AI Ethics and Ethics of Vulnerability

Kerstin Schlögl-Flierl, Paula Ziethmann

Efficiently address risks, such as allocating welfare delivery resources to those in need, requires accurate risk assessments and predictions in the field of social work. (cf. Schrödter, Bastian, and Taylor 2018) Through various factors, such as AI's ability to meet these challenges accurately (risk assessments and predictions) but also through an increasing shortage of staff in social work, there is optimism, that AI can improve and facilitate processes in this field.

The aim of this paper is to bring a critical position into the debate about the use of AI in child welfare. However, the focus should not be on a false faith in technology. (cf. Klöcker 2021) The human judgement is not unchallenged, so we are not arguing here that decisions should exclusively be made by humans. (cf. Schwabe 2022) For example, a study supported by the Federal Anti-Discrimination Agency has stressed, with regard to learning algorithms in particular, that in many cases human decisions can be the sources of discrimination risks. (cf. Orwat 2019)

As in any ethical judgment, the status quo analysis must first be presented. We will do this based on an algorithm in New Zealand and the situation in Germany (chapter 2). In this paper, we will then identify the thresholds that must be considered when using AI in any way in this field. In our opinion, the aspects of social work and ethics should not be separated, and the stakeholder's perspective is necessary. That is why an Ethics of Vulnerability forms the basis in chapter 3. An Ethics of Vulnerability has a social ethical aspect (the community exacerbating vulnerability or changing structures to reduce vulnerability) in addition to the individual ethical aspect (the individual as vulnerable). In assessing the best interests of the child, all parties involved are vulnerable, and thus, we present several precautionary reasons in chapter 4 that should be con-

sidered. Finally, we argue that our innovative approach, which combines the Ethics of Vulnerability with the Ethics of AI, is the best way to meet the needs and demands in child welfare.

1. Status Quo Analysis

To conduct an ethical assessment, it is necessary to begin with an analysis of the status quo. For this, we will first focus on the so-called Vulnerable Children PRM in New Zealand.

The philosopher Tim Dare reports:

> »The Vulnerable Children PRM […] was developed and validated using an anonymized dataset linking administrative records from New Zealand's welfare benefit and Child, Youth and Family Services system for children who were born between January 2003 and June 2006 and had a benefit spell before the age of two: a sample of 57,986 children comprising about 33 per cent of all children born in New Zealand during that period. The PRM algorithm was developed by identifying potential variables in cases of substantiated maltreatment in 70 per cent of the sample. 132 variables – including demographic and historical features of a child, their family, household and community – were found to make a statistically significant contribution to the model and were therefore retained in a ›core algorithm‹ which was tested on the remaining 30 per cent validation sample. The algorithm generated a risk decile score at the start of each new benefit spell for each child in the sample with 10 indicating a child as being within the top 10 per cent of risk, down to 1 as being in the bottom 10 per cent.« (Dare 2015, 65)

Concerning children with risk scores of 9 and 10 suggested an intervention. That means intensive intervention concerning 5 per cent of the total population. The PRM eventually was not used because of serious concerns, which are explained below.

According to Philip Gillingham, an Associate Professor, ARC Future Fellow and Associate Investigator at ARC Centre of Excellence in Automated Decision-Making and Society at University of Queensland, the Vulnerable Children PRM in New Zealand has serious flaws. Gillingham notes that the algorithm used by the system associated child abuse with the period of social assistance received, single parenthood, and previous contact with child protective services. (cf. Gillingham 2021)

Two of these are more than problematic in Gillingham's eyes:

> »First, these predictors occur in the average population and thus do not distinguish the group of children who are at increased risk with respect to child maltreatment. Most single parents on welfare do not maltreat their children.« (Gillingham 2021, 32)

Secondly, these predictors are indicators of poverty. Most of the poor parents do not maltreat their children.

> »This is not to deny that poverty may be associated with child maltreatment. What is problematic is the indispensability of these predictors in the case of the data used in the training of the algorithm (as mentioned above, but especially the data of welfare receipt) and its implementation (the screening of families entering the welfare system).« (Gillingham 2021, 32)

Thirdly, the use of the indicator »previous contact with child protective services« causes different questions: Does this mean that the child protective service has not been effective? Are reports made to child protective services but not investigated further after initial screening included? In the case of the child that had not suffered abuse or was at risk of abuse: can the algorithm differentiate between these reports?

In case of the New Zealand Report, Gillingham resumes, all contacts with child protective services are considered to be against the parents and increase the likelihood that they will be re-investigated. (cf. Gillingham 2021)

To improve the accuracy of risk assessments and predictions, it is important to reduce the false-positive rate, which refers to the incorrect classification of families as high risk. One potential solution is to choose higher thresholds for intervention. Dare suggests among others:

> »Providing opportunity for experienced child protection professionals to exercise judgement about appropriate responses to a family's identification as at risk. Ensuring that such professionals understand the potential of the Vulnerable Children PRM to miscategorise families. Ensuring that intervention triggered by identification as at risk is as non-intrusive as possible, consistent with the overall aims of reducing child maltreatment risk.« (Dare 2015, 73)

These are first ideas for precautionary thresholds.

In **Germany**, an understanding of social work in the field of child welfare can be increased by examining relevant laws. According to § 8a SGB VIII, an assessment is legally required in cases where a child's well-being is at risk:

Protection mandate in the event of a risk to the well-being of a child:[1]

> »If the youth welfare office becomes aware of serious indications that the welfare of a child or adolescent is at risk, it shall assess the risk of danger in cooperation with several specialists. Insofar as the effective protection of this child or adolescent is not in question, the youth welfare office shall involve the legal guardians as well as the child or adolescent in the risk assessment. And, this is necessary according to professional judgment, to get a direct impression of the child and his or her personal environment. As well as to involve persons who have transmitted data to the youth welfare office in accordance with Section 4 (3) of the Act on Cooperation and Information in Child Protection in the risk assessment in an appropriate manner.«

That means: in the German child protection, a ›security/safety (as subjective feeling) orientation‹ dominates: in particular, events of present or past harm to the child's well-being are to be assessed concerning socio-educational need for action.[2] (cf. Bastian, Freres, and Schrödter 2017)

Professionalism is also a crucial aspect of the system, requiring the use of tacit knowledge gained through experience as a social worker. The ability to observe (odors for example) and exercise undirected attention is essential, and strong observational skills are necessary. The process is more geared towards creating an impression of a situation, leading to open dialogues with critical reconstructive potential, rather than a rigid subsumption logic.

As such, the introduction of an algorithm into this field must be approached with caution, especially in the field of child and family welfare.

In ethical terms, you must admit in a first step that child welfare is less objective than expected (cf. Wiesemann 2016). It is crucial to acknowledge the complex nature of child welfare and the competing rights at play, including the parental right to upbringing, the child's right to privacy, and the child's right to protection (cf. Gutwald et al. 2021). In Germany, parents or guardians have the fundamental freedom and right to educate their children according to their

1 Translation by Kerstin Schlögl-Flierl.

2 However, if you have a further look on the algorithm in the US, especially the Allegheny one or the one in New Zealand: One dominant predictor is harm. So the risk score also are based on the ›harm orientation‹, but are more prognostic.

own ideas within the framework of a private sphere, without the state intervening. This is therefore a defensive right against the state. The state may only intervene in this sphere of freedom if a child's physical, mental or psychological well-being is at risk and the parents are not (or no longer) willing or able to counteract this risk. From an ethical point of view, this raises the further question of what rights children themselves have under Article 16 of the UN Convention on the Rights of the Child to a private life in their family, and to what extent this right may conflict with their right to protection. Professionals in child protection must carefully consider these rights and the specific circumstances of each case.

The concrete situation is to be examined once again in detail. What are the circumstances in this decision-making?

> »Only rarely does an assessment of an acute danger to the life and limb of a child occur in child protection, and there is not always a consensus among those involved as to whether and in what form a child is being neglected or abused. Rather, child protection professionals have to deal with complex life situations of children and adolescents and their families and with multifactorial causal and contextual conditions for child abuse and neglect, where clarity is often lacking and the scope for interpretation is large.« (cf. Gedik and Wolff 2021, 418) [3]

The intricacies of child welfare to grasp fully, it is crucial to acknowledge the competing rights at play, including the parental right to upbringing, the child's right to privacy, and the child's right to protection. However, it is not always clear if and in what form a child is being neglected or abused, and consensus may not always be reached among those involved. Therefore, we propose addressing this situation with an Ethics of Vulnerability, which recognizes the complex and multifaceted nature of child welfare and the need to address the needs and demands of all parties involved. This approach emphasizes the importance of acknowledging and responding to the vulnerabilities of children and families in the child welfare system, as well as the vulnerabilities of the professionals involved in child protection work. By taking a holistic approach and acknowledging the various dimensions of vulnerability, we can better address the complexities of child welfare and work towards the best interests of

3 Translation by Kerstin Schlögl-Flierl.

the child, his or her parents, and the family as a whole. (cf. Gedik and Wolff 2021, 419)

2. Ethics of Vulnerability

A more theological, but not exclusive (cf. Keul 2021, Quast-Neulinger 2018), approach in this field could be to speak of an Ethics of Vulnerability. Vulnerability become negativ connoted by political talk in the context of the Corona pandemic. (cf. Schmidhuber 2020) The most vulnerable groups ranked high on the vaccine prioritization list. Nevertheless, by what and in what ways does one belong to the vulnerable in the context of child welfare in social work?

In the scientific discussion of vulnerability (cf. Mackenzie 2014) in various disciplines, vulnerability certainly been formatted as a risk, but at the same time, it can also be seen as a resource or an asset, as will be unfolded below. Fundamental are the following distinctions: every person is vulnerable (ontological vulnerability), some by the situation moreover (situational vulnerability, e.g. imagine a family with a small flat in the pandemic). To be distinguished from this is vulnerability generated by structure. This would be, for example, the nursing home residents, who ›suffered‹ not a little from the various regulations in the pandemic.

Because all people are to be understood as fundamentally vulnerable, the risk is less in the center of thinking, or should be for ethics, but more also the positive direction of an understanding of the human being as a vulnerable being. (cf. Haker 2021) The openness to others and to oneself this entails is not to be underestimated. An Ethics of Vulnerability has besides the individual-ethical cut (the individual as the vulnerable) also a social-ethical (the community, which aggravates vulnerability or changes structures to reduce vulnerabilities) aspect.

In the child welfare assessment, all stakeholders are vulnerable. First the child or children, then the family in its relationship, the social worker in his or her very difficult task, the state as protector/advocacy of family and children and concerning the use of AI. The designers and programmers who develop the AI that will be used. This certainly is not comprehensive. That is why different precautionary reasons should be concerned, which we will present in the next chapter.

3. Precautionary reasons

An Ethics of Vulnerability implies that not only the children or families are vulnerable but also the social worker. Schrödter, Bastian and Taylor prognoses:

> »In the future, it will no longer be a question of the effectiveness of statistical forecasting methods, nor of whether social work should allow statistical methods to enter practice at all. In the long term, this process will probably be unstoppable. Therefore, it must increasingly be about the ethical questions of what consequences these powerful procedures will have for social work practice and how the negative consequences can be legally and technically tamed […].« (Schrödter, Bastian, and Taylor 2020, 256)[4]

In this part, no distinction is made between the different algorithms, and data protection is also excluded from the considerations, which, however, must be dealt with in general and also for this field. The following concerns relate mainly to the various vulnerabilities.

3.1 Child/Family: Stigmatization of already vulnerable populations

Stigmatization of already vulnerable groups is a pervasive issue in the use of AI. (cf. Schneider and Seelmeyer 2018, 24) However, in the field of child welfare, it means increasing pressure on perhaps already struggling families and households.

Sometimes even the term itself can be misleading: ›High risk‹. A better choice here would be to speak of a high priority for services for the struggling families. It should not happen that the consequence of using AI is a reduced readiness to engage with service providers and leading professionals to deal differently with stigmatized individuals. (cf. Dare 2015, 69) Such effects would be particularly problematic for child protection, as this would affect an already highly stigmatized group, who would then be even more subject to suspicion and control.

Moreover, insofar as the goal of big data analytics is to identify at-risk populations, this promotes an individualistic conception of »risk« at the expense of more justice-oriented conceptions: (cf. Bastian et al. 2018) Racial disparity is an often mentioned danger in the use of AI. (cf. Dare and Gambrelli 2017,

4 Translation by Kerstin Schlöfl-Flierl.

5) Poor people, often intersectional discriminated against by the category of race, are also stigmatized. Thus, the individualistic concept of »risk« leads to a dangerous neo-liberal view those vulnerable populations are solely responsible for their situation and distracts from systemic and institutional problems that would be addressed by justice oriented approaches.

Furthermore, the use of AI can result in the collection of information about family members beyond the child in question. This can also be grasped under the name »dataveillance«, so called social surveillance through digitalization. Thus, forecasting can more effectively organize the exclusion of those deemed dangerous, criminal, in need of help, or otherwise deviant, even though these populations constructed as such by surveillance technologies in the first place. (cf. Bastian et al. 2018)

3.2 Social Worker: The challenged role

Imagine the case, the social worker decides against the recommendation of the algorithm and then happens the case of child endangerment. »This example illustrates the new tension between human decision-making authority and ›algorithmic authority‹ in the media public sphere.« (cf. Gapski 2020) One should add not only, but also in the media public.

This concerns the understanding of professionalism. It is therefore unclear what value intuition has in the judgment and action processes of social work and what value feelings should have for professionalized action practice. (cf. Bastian 2018, 128f.) A brief look at the history of social work already shows that standardization by means of legal requirements has also found its way into this area. This is associated with an immediate safeguarding of decisions (cf. the risk assessment forms, Dare 2015). On the one hand, it may be argued that the problem of standardization could be an increasing decline in reflexive professionalism. In addition, a decoupling of decision and responsibility must be counteracted already without Artificial Intelligence. On the other hand, the potential can be seen that for social workers, ethically based digital tools for the assessment of child endangerment could lead to a significant relief, as it is formulated in the description of KAIMo. (cf. Homepage; cf. Bastian 2012) However, this requires both ›data literacy‹ to handle complex data sources and sophisticated procedures for processing them, as well as consistent privacy protection for the use of diverse data sources as new requirements for the professionalism of specialists. (cf. Schneider and Seelmeyer 2018, 24) Furthermore, possible de-skilling must be actively counteracted.

These issues at hand are multifaceted, and it is crucial to identify which of the several problems can be addressed by the implementation of AI? In the field of child welfare, Gedik and Wolff have outlined the following complexities:

> »Such an investigation and understanding takes time, professional knowledge and skills. What is needed above all, however, are professional specialists with an attitude of solidarity who are able to make contact with families, parents and children in need, to meet them courageously and to talk to them in a trusting and non-judgmental manner about sensitive questions of life and upbringing, about the relationships they have developed with each other, possibly about painful life stories or also about debts and other material and social burdens. The professionals must have methodical competence to examine the conflict and problem situation with the families, to understand it well and to bundle it in a multi-perspective way. And they must be linguistically capable of openly and comprehensibly presenting the necessary assistance to the parties involved from a professional perspective as well, to justify, explain and negotiate well with special attention to the child involved, his personality and health, his rights, needs, wishes and interests, but without neglecting his parents and the entire family.« (Gedik and Wolff 2021, 420)[5]

3.3 The profession: The challenged role of social work, in general

In this section, we shift our focus from the specific field to a broader perspective on social work. There is an ongoing debate in the field of social work regarding mandates. In a classical approach, two mandates have been identified: one of the client and one of the state. Unsurprisingly, there is potential for conflict here – and presumably, the state will view the use of AI differently than clients. However, there is also a third mandate that has emerged in recent times – the mandate of social work as a profession, in the words of Silvia Staub-Bernasconi: a mandate for human rights. (cf. Staub-Bernasconi 2019) There has been a historical shift in the way social work considers itself as a profession. It once was seen as an act of mercy, then as a service, and now as a human rights profession.

By understanding social work as a human rights profession, it becomes possible to incorporate the view of human rights: How are children's rights and family law respected. One must distinguish between at least three levels. Social

5 Translation by Kerstin Schlögl-Flierl.

work at the micro level: in collaboration with the client; social work at the meso level: in the affiliation with social institutions; and social work at the macro level: in the responsibility for the society. (cf. Leith 2021, 330ff.)

On which level AI can be supporting? The idea also presupposes that the ethical decision can be transferred into algorithmic structures at all. This may be possible for strictly deontic or utilitarian ethics. In the eyes of Björn Görder, however, the ethics of social work in particular usually ties in with other ethical traditions (at least also) for good reasons. These include approaches based on virtue, situational and care ethics, which take into account attitudes, virtues, motives, relationship quality, feelings and also distress, or moral intuitions. Values such as classically spoken charity, empathy, loyalty, commitment cannot be translated into algorithms and presuppose knowledge of the world that is relevant to interpersonal encounters. (cf. Görder 2021)

Besides these problems of what ethics theory is binding, questions of justice also arise. These are the typical resource allocation issues (cf. Dare 2015, 64) but can be reinforced by the use of AI. It begins by the payment of social workers:

> »It might be difficult, for instance, to maintain the important existing relationships between families and the nurses and community health workers who deliver current universal services while changing their focus and intensity: it may matter that the existing services are relatively light-touch. Further, universal programmes are expensive, and they would be even more so if made more intensive.« (Dare 2015, 69)

With the introduction of AI, it is important not to dismantle other social work services but to make them more targeted and effective.

4. Relationship between Ethics of Vulnerability and AI Ethics

A vulnerability ethics approach stresses the perspective of the needs and requirements of all involved stakeholders: the need of the child, the need of the family and the need of the social worker (and the Advocacy of the State); the social worker who wants to have more security/more ›hedging‹ for his or her decision; the need of the family: prevention, support and privacy and the need of the child: in general to have a good future. It is important to note that the algorithm is just one element in the field of decision-making in social work.

AI Ethics revolves around sociotechnical change and the impact on society and the lives of individuals. Because AI Ethics addresses both individual ethical and social ethical perspectives, it fits very well with the demonstrated discussion of the use of AI in child welfare. By combining AI Ethics and Ethics of Vulnerability, we will be able to show vulnerability of stakeholders and the impact of AI on the individual and systemic situation, which will allow us to address the needs and requirements of stakeholders in the best possible way.

To this end, in this chapter we will first examine the AI ethical implications of using AI in child vulnerability, as demonstrated in Cheng's et al. new empirical study (cf. Cheng et al. 2022), and then integrate it with our Vulnerability Ethics Approach. The aforementioned study analyzes the decision-making of child welfare call screening worker over a span of four years, in Allegheny County, Pennsylvania. The workers use the AI-based Allegheny Family Screening Tool (AFST) to decide about which reports of child abuse or neglect (henceforth referrals) to investigate. For example, the conduct interviews in which they ask the workers how they incorporate algorithmic predictions into their decision-making process. When people work with algorithms in a child welfare context as is known to have racial disparities, will they serve to mitigate or exacerbate disparities? The answer to this question can inform the responsible design and use of AI tools in the child welfare context, as well as other high-stakes social decision-making contexts. The study showed that, compared to the algorithm alone, workers reduced the disparity in screen-in rate between black and white children from 20 % to 9 %. Cheng et al. showed that workers achieved this by making holistic risk assessments and adjusting for the algorithm's limitations.

The authors call for the promotion of more collaborative decision-making. They argue that the fact that child welfare caseworkers often made decisions collaboratively led to, among other things, a reduction in individual bias. However, when working together formally, caseworkers would feel they had little influence – even when making screening recommendations. Thus, Cheng et al. advocate encouraging more regular conversations between caseworkers and supervisors about caseworker recommendations so that they are not overlooked. In this regard, they argue that the AFST interface could also provide opportunities to foster informal collaboration. For example, future versions of AFST or similar tools could include a feature that suggests clerks who have handled similar referrals in the past. This could allow workers to collaborate with the right person for each referral, which could be particularly helpful for very uncertain referrals.

According to the study, child welfare workers were generally aware of their own individual biases. Another way to curb this bias, according to the authors, could be to increase diversity among child welfare staff, particularly among supervisors who make the final decisions.

In evaluating the study, we will draw closely on our findings about the Ethics of Vulnerability. We will first address the needs and demands of the social worker, then those of the family, and finally those of the children.

Workers' need for more security

Regarding the needs of social workers to be more secure in their decision-making, AI could be used as a reflective tool. Here, the AI ethical aspects of explainability and transparency must be addressed primarily. It is crucial for social workers to understand the limitations of AI, such as potential biases and what the AI cannot know compared to them. Only with this understanding AI can be used as a decision support tool alongside with other social workers and supervisors to assist the individual social worker.

Family's need for prevention, support, privacy

In order to ensure that marginalized groups are not further disadvantaged by the use of AI, so that appropriate families can receive support, ethical considerations of anti-bias or rather bias awareness, as well as the avoidance of dataveillance, must be taken into account. We suggest that the terminology used should acknowledge that there is no completely bias-free AI, and therefore »bias awareness« must be prioritized. If social workers are aware of the limitations of the AI tool, as mentioned in the previous point, bias awareness can be increased, and discrimination reduced. Additionally, as mentioned in section 3.1, we propose replacing »high risk« with »high priority« to avoid perpetuating biases. To protect the privacy of families and prevent dataveillance, a high level of data protection must be maintained.

To address not only individual families but also systematic and institutional issues, AI could be utilized to analyze data for intersectionality. As our previous analysis of the status quo in chapter 1. has demonstrated, various parameters appear to be interconnected, such as poverty and single parenthood. The question then arises: how can the social work system acknowledge and support this interconnectivity without perpetuating any biases or prejudices?

The needs of the children

To address this need of the child is the most challenging and involves the most considerations. Depending on the child's level of maturity, their decision about whether they want to stay with their family can not be taken into account. This decision relies heavily on the experience of the social workers: do they believe that the child is able to express their will accurately, or are they under pressure to stay in their current situation, and therefore not expressing their true desires? The recommendation of the AI must be understood in its entirety, including all its limitations, so that it can be incorporated into the decision-making process. Although AI can make good predictions, it cannot take individual cases into account and can thus exacerbate the child's situation.

Summary

The AI is helpful because of velocity, more efficiency, and more precision. This is undeniable the balancing between different needs and interests in this field. (cf. Gapski 2020) The social worker can get helpful hints by the algorithm but his and her professionalism should not be touched and his and her ›hedging‹ of decisions that are by no means easy.

This paper critically examined the use of AI in child welfare. We argued that the social work and ethics aspects should not be separated, but rather considered from the perspective of all stakeholders. The paper presented a status quo analysis based on an algorithm in New Zealand and the situation in Germany, and identified thresholds that needed to be considered when using AI in this field. The Ethics of Vulnerability was presented as a basis for weighing the best interests of the child, which included both social and individual ethical aspects. In the next chapter, resulting precautionary grounds were identified that should be taken into account. We highlighted that our innovative approach of combining the Ethics of Vulnerability with the Ethics of AI was the best way to address the needs and demands in child welfare. We highlighted the potential benefits and ethical considerations associated with the use of AI in child welfare and protection. The needs and demands of social workers, families and children were each addressed through an Ethics of Vulnerability, and the importance of simultaneously considering AI ethical issues such as transparency, bias awareness and data security was emphasised. In addition, it was suggested that AI could be used to examine systemic and institutional issues

by analysing data for intersectionality. While AI could increase efficiency, accuracy and speed, these benefits needed to be balanced with the professionalism and decision-making skills of social workers. Ultimately, the use of AI in child welfare had to be done with caution, careful consideration and adherence to ethical principles.

References

Bastian, Pascal, Mark Schrödter, Roland Becker-Lenz, Joel Gautschi, Martin Grosse, Martin Hunold, and Cornelia Rüegger. 2018. »Bauchgefühle in der Sozialen Arbeit«. In *Wa(h)re Gefühle. Sozialpädagogische Emotionsarbeit im wohlfahrstaatlichen Kontext*, edited by Kommission Sozialpädagogik, 128–140. Weinheim: Beltz Juventa.

Bastian, Pascal. 2012. »Die Überlegenheit statistischer Urteilsbildung im Kinderschutz. Plädoyer für einen Perspektivwechsel hin zu einer angemessenen Form sozialpädagogischer Diagnosen.« In *Rationalitäten des Kinderschutzes*, edited by Thomas Marthaler, Pascal Bastian, Ingo Bode, and Mark Schrödter, 251–271. Wiesbaden: Springer VS. DOI: https://doi.org/10.1007/978-3-531-19146-1_11.

Cheng, Hao-Fei, Logan Stapleton, Anna Kawakami, Venkatesh Sivaraman, Yanghuidi Cheng, Diana Qing, Adam Perer, Kenneth Holstein, Zhiwei Steven Wu, and Haiyi Zhu. 2022. »How Child Welfare Workers Reduce Racial Disparities in Algorithmic Decisions«. *CHI Conference on Human Factors in Computing Systems* (CHI '22), April 29-May 5, 2022. New Orleans, LA, USA, 1–22. DOI: https://doi.org/10.1145/3491102.3501831.

Dare, Tim, Janice McGhee, Brigid Daniel, Andrew Cooper, Kay Tisdall, Jason Hart, Trevor Spratt, Tarja Pösö, Fiona Arney, Stewart McDougall, et al. 2015. »The Ethics of Predictive Risk Modelling«. In *Challenging Child Protection: New Directions in Safeguarding Children*, edited by Lorraine Waterhouse, 64–76. London: Jessica Kingsley Publishers.

Dare, Tim, and Eileen Gambrill. April 2017. *Ethical Analysis: Predictive Risk Models at Call Screening for Allegheny County*. Accessed March 16, 2023. https://www.alleghenycountyanalytics.us/wp-content/uploads/2019/05/Ethical-Analysis-16-ACDHS-26_PredictiveRisk_Package_050119_FINAL-2.pdf.

Gapski, Harald. 2020. »Digitale Transformation. Datafizierung und Algorithmisierung von Lebens- und Arbeitswelten«. In *Handbuch Soziale Arbeit und Digitalisierung*, edited by Nadia Kutscher, Thomas Ley, Udo Seelmeyer, Frie-

derike Siller, Angela Tillmann, and Isabel Zorn, 156–166. Weinheim: Beltz Juventa.

Gedik, Kira, and Reinhart Wolff. 2021. »Kindesmisshandlung und Vernachlässigung untersuchen: Gefährdungen einschätzen«. In *Kinderschutz in der Demokratie – Eckpfeiler guter Fachpraxis. Ein Handbuch*, edited by Kira Gedik, and Reinhart Wolff, 415–451. Opladen, Berlin, Toronto: Barbara Budrich Verlag.

Gillingham, Philip. 2021. »Big Data, Prädiktive Analytik und Soziale Arbeit«. In *Sozial Extra* 1: 31–35.

Görder, Björn. 2021. »Die Macht der Muster. Die Ethik der Sozialen Arbeit vor professionsbezogenen und gesellschaftlichen Herausforderungen durch ›künstliche Intelligen‹«. In *Ethik Journal* 7 (2), Heft 1, Download unter: https://www.ethikjournal.de/fileadmin/user_upload/ethikjournal/Texte_Ausgabe_2021_2/Goerder_Ethikjournal_2.2021.pdf (Zugriff am 05.02.2024).

Gutwald, Rebecca, Jennifer Burghardt, Maximilian Kraus, Michael Reder, Robert Lehmann, and Nicholas Müller. 2021. »Soziale Konflikte und Digitalisierung. Chancen und Risiken digitaler Technologien bei der Einschätzung von Kindeswohlgefährdungen«. In *Ethik Journal* 7 (2), Heft 1, Download unter: https://www.ethikjournal.de/fileadmin/user_upload/ethikjournal/Texte_Ausgabe_2021_2/Gutwald_u.a._Ethikjournal_2.2021.pdf (Zugriff am 05.02.2024).

Haker, Hille. 2021. »Verletzliche Freiheit. Zu einem neuen Prinzip der Bioethik.« In *Theologische Vulnerabilitätsforschung: Gesellschaftsrelevant und Interdisziplinär*, edited by Hildegund Keul, 99–118. Stuttgart: Kohlhammer.

Keul, Hildegund. 2021. *Schöpfung durch Verlust*, Band 1. Würzburg: Echter Verlag.

Keul, Hildegund. 2021. *Schöpfung durch Verlust. Eine Inkarnationstheologie der Vulnerabilität, Vulneranz und Selbstverschwendung*, Band 2. Würzburg: Echter Verlag.

Klöcker, Katharina. 2021. »Autoritäre Algorithmen: Wenn Künstliche Intelligenz Entscheidungen trifft«. In *Herder Korrespondenz* 75 (53): 37–39.

Leith, Katherine. 2021. *Grundlagen ethischen Handelns in der Sozialen Arbeit*. 2. korrigierte und aktualisierte Auflage. Bremen: Apollon University Press.

Mackenzie, Catriona, Wendy Rogers, and Susan Dodds. 2014. »Introduction: What is Vulnerability and Why Does It Matter for Moral Theory?« In *Vulnerability: New Essays in Ethics and Feminist Philosophy*, edited by Catriona Mackenzie, Wendy Rogers, and Susan Dodds, 1–20. Oxford: Oxford University Press.

Neulinger, Michaela. 2018. *Zwischen Dolorismus und Perfektionismus. Konturen einer politischen Theologie der Verwundbarkeit*. Paderborn: Ferdinand Schöningh.

Orwat, Carsten. 2019. *Diskriminierungsrisiken durch Verwendung von Algorithmen*. Berlin: Accessed March 16, 2023. https://www.antidiskriminierungsstelle.de/SharedDocs/downloads/DE/publikationen/Expertisen/studie_diskriminierungsrisiken_durch_verwendung_von_algorithmen.pdf;jsessionid=ED36831F7BC054A83DFF38D25EA5D6B2.intranet242?__blob=publicationFile&v=3.

Pascal, Bastian, Katharina Freres and Mark Schrödter. 2017. »Risiko und Sicherheit als Orientierung im Kinderschutz. Deutschland und USA im Vergleich«. In *Soziale Passagen* 9: 245–261.

Schmidhuber, Martina. 2020. »Vulnerabilität in der Krise«. In *Die Corona-Pandemie. Ethische, gesellschaftliche und theologische Reflexionen einer Krise*, edited by Wolfgang Kröll, Johann Platzer, Hans-Walter Ruckenbauer, Walter Schaupp, 271–282. Baden-Baden: Nomos. DOI: https://doi.org/10.5771/9783748910589.

Schneider, Diana, and Udo Seelmeyer. 2018. »Der Einfluss der Algorithmen. Neue Qualitäten durch Big Data Analytics und Künstliche Intelligenz«. In *Sozial Extra* 3: 21–24.

Schrödter, Mark, Pascal Bastian, and Brian Taylor. 2018. »Risikodiagnostik in der Sozialen Arbeit an der Schwelle zum ›digitalen Zeitalter‹ von Big Data Analytics.« *ResearchGate*. Accessed March 16, 2023. https://www.researchgate.net/publication/328216929_Risikodiagnostik_in_der_Sozialen_Arbeit_an_der_Schwelle_zum_digitalen_Zeitalter_von_Big_Data_Analytics.

Schrödter, Mark, Pascal Bastian, and Brian Taylor. 2020. »Risikodiagnostik und Big Data Analytics in der Sozialen Arbeit«. In *Handbuch Soziale Arbeit und Digitalisierung*, edited by Nadia Kutscher and Thomas Ley et al., 255–264. Weinheim: Beltz Juventa.

Schwabe, Matthias. 2022. *Die »dunklen Seiten« der Sozialpädagogik. Über den Umgang mit Fehler, Unvermögen, Ungewissheit, Ambivalenzen, Idealen und Destruktivität*. Weinheim: Beltz Juventa.

Staub-Bernasconi, Silvia. 2019. *Menschenwürde – Menschenrechte – Soziale Arbeit. Die Menschenrechte vom Kopf auf die Füße stellen*. Opladen, Berlin, Toronto: Barbara Budrich Verlag.

Wiesemann, Claudia. 2016. »Kindeswohl – Ein Problemaufriss aus der Perspektive der Medizinethik.« In *Zeitschrift für medizinische Ethik* 62: 235–244.

Verantwortungsvolle Empfehlungssysteme für die medizinische Diagnostik

Daniel Schlör, Andreas Hotho

1. Einleitung

Die frühe Entwicklung und Einführung von Krankenhaus- und Gesundheitsinformationssystemen hat die fortschreitende Digitalisierung von Prozessen in Krankenhäusern gefördert. Viele dieser Prozesse, die früher Schriftverkehr und telefonische Absprachen erforderten, sind heute in IT-Lösungen integriert und erfordern die Interaktion von Ärzt*innen und medizinischem Personal mit entsprechenden Schnittstellen und Tools. Diese Umstellung auf digitales Datenmanagement und Prozessunterstützung kommt der Versorgung der Patient*innen zwar in vielerlei Hinsicht zugute, erfordert aber von den Ärzt*innen eine genaue digitale Erfassung aller relevanten Informationen für Abrechnungs- und Dokumentationszwecke und damit oft eine Auswahl aus einer unüberschaubar großen Menge an Möglichkeiten. Das beansprucht viel Zeit, die sonst für die ärztliche Versorgung der Patient*innen aufgewendet werden könnte. Die systematische Erfassung von Gesundheitsdaten über einen langen Zeitraum hinweg bietet jedoch Möglichkeiten, diesen Prozess zu verbessern und das medizinische Personal durch die Einführung von Empfehlungssystemen zu unterstützen.

Betrachten wir zum Beispiel eine Ärztin in einem Krankenhaus, die eine Untersuchung bei einem Patienten durch eine andere Abteilung durchführen lassen möchte. Dazu muss sie eine entsprechende Anfrage stellen, die die Anamnese und die medizinische Fragestellung enthält, und die richtige Untersuchung auswählen. Diese Auswahl aus einer großen Menge von Untersuchungen kann durch ein Empfehlungssystem unterstützt werden, das entweder inhaltsbasiert arbeitet, oder auf ähnlichen Fällen aus der Vergangenheit basiert. Inhaltsbasierte Empfehlungen könnten in diesem Beispiel etwa

Anamnese, medizinische Fragestellung, mögliche medizinische Vorgeschichte des Patienten sowie anderes medizinisches Wissen berücksichtigen, um eine möglichst passende Empfehlung für die Untersuchung auszusprechen. Für Empfehlungen anhand ähnlicher Fälle würden beispielsweise hinsichtlich der vorherigen Behandlungssequenz oder Patientenzustand ähnliche Fälle betrachtet werden und die in diesen Fällen durchgeführten Untersuchungen empfohlen werden.

Beide Ansätze haben technische Vor- und Nachteile, die in Abschnitt 2.1 genauer benannt werden, sind aber auch für unterschiedliche Anwendungszwecke verschieden gut geeignet. Betrachtet man beispielsweise speziell radiologische Untersuchungen, was genau den oben genannten Fall zwischen anfordernder und erfüllender Abteilung darstellt, zeigen sich große Unterschiede, je nach anfordernder Stelle: Während in der Onkologie beispielsweise häufig ähnlich gelagerte Fälle betrachtet werden, die dementsprechend auch gut durch Empfehlungen basierend auf ähnlichen Fällen unterstützt werden können, finden sich in der Notaufnahme oder Unfallchirurgie sehr unterschiedlich gelagerte Fälle, die eine einzelne Betrachtung basierend auf der konkreten Anamnese erfordern, insbesondere auch, weil häufig zunächst keine Krankengeschichte der Patient*in verfügbar ist.

Mit diesen verschiedenen Perspektiven eröffnen sich aber auch unterschiedliche ethische Fragen für den Einsatz von Empfehlungssystemen in diesem sensiblen medizinischen Kontext, die berücksichtigt werden müssen, um ein verantwortungsvolles Verhalten des Systems selbst sowie in der Interaktion mit dem medizinischen Personal sowie den Patient*innen zu gewährleisten. Diese Überlegungen betreffen beispielsweise das Krankenhauspersonal, etwa wenn es um deren Effizienz, mögliche Kosten- und Zeitersparnisse und den Konsequenzen für die Arbeitsplätze geht, aber auch die Patient*innen, die davon profitieren könnten. Fragestellungen hinsichtlich der Datenqualität und möglichen beispielsweise diskriminierenden Verzerrungen, die ein solches System verstärken könnte, aber auch Überlegungen, wer im Kontext eines solchen sozio-technologischen Systems welche Verantwortungen und damit Rechenschaftspflichten hat, sind hier relevant.

Ausgehend von diesen Überlegungen werden in diesem Beitrag Kriterien für ein verantwortungsbewusstes Empfehlungssystem im medizinischen Kontext aus einer anwendungsorientierten Perspektive skizziert und mögliche Designentscheidungen mit besonderem Fokus auf die bereits genannten Fragen sowie Sicherheit, Transparenz und Konformität hinsichtlich der anzuwendenden Regularien betrachtet. Hierfür nehmen wir aus der informa-

tischen Perspektive den Standpunkt für ein konkretes Anwendungsszenario ein und werfen dafür relevante Fragestellungen auf, die von den technischen Aspekten losgelöst aus ethischer Sicht betrachtet werden können.

2. Grundlagen

In diesem Kapitel sollen zunächst die Grundlagen zu Empfehlungssystemen sowie die hier betrachtete Operationalisierung ethischer Aspekte vorgestellt werden.

2.1 Empfehlungssysteme

Empfehlungssysteme sind computergestützte Algorithmen und Systeme, die Anwender*innen auf Grundlage ihrer Vorlieben, Interessen oder ihrem bisherigen Verhalten Artikel oder andere Objekte vorschlagen (Bobadilla et al. 2013). Diese stellen im klassischen Sinne beispielsweise Produkte in Online-Shops (Fischer, Zoller und Hotho 2021), Bücher (Mooney und Roy 2000), Musik (Song, Dixon und Pearce 2012) und Videos (Gomez-Uribe und Hunt 2015), aber auch weniger typische Objekte, wie beispielsweise Schlagworte (Jäschke et al. 2007), Gegenstände in Computerspielen (Dallmann et al. 2021), medizinische Behandlungen (Sun et al. 2016) oder in unserem Beispiel medizinische Untersuchungen dar und sollen im Folgenden unter dem Begriff *Artikel* subsumiert werden. Empfehlungssysteme spielen eine entscheidende Rolle bei der Unterstützung der Anwender*innen z.B. bei der Navigation durch die großen Mengen an verfügbaren Informationen und verbessern so den Entscheidungs- oder Auswahlprozess (Zhang et al. 2019).

Empfehlungssysteme nutzen dafür Algorithmen und Techniken, um Daten der Nutzenden zu analysieren und häufig personalisierte Empfehlungen zu generieren. Durch die Berücksichtigung von Benutzungsfeedback (Shi, Larson und Hanjalic 2014) oder vorheriger Interaktionen (Adomavicius und Tuzhilin 2005) zielen solche Systeme darauf ab, den Anwender*innen relevante und hilfreiche Vorschläge zu unterbreiten. Der Hauptzweck von Empfehlungssystemen besteht darin, bei der Entdeckung neuer Artikel und Ressourcen unter anderem bezüglich der persönlichen Präferenzen zu unterstützen (Zhang et al. 2019), indem sie die sich teilweise entgegenstehenden Faktoren wie Genauigkeit, Neuartigkeit, Streuung und Stabilität der Empfehlungen balancieren (Bobadilla et al. 2013).

Hinsichtlich der verwendeten Methoden lassen sich Empfehlungssysteme in drei größere Kategorien einteilen: Systeme, die auf Collaborative Filtering (Adomavicius und Tuzhilin 2005) basieren, solche, die Content-based Filtering (Pazzani und Billsus 2007) nutzen, sowie Hybride Ansätze (Balabanović und Shoham 1997).

2.1.1 Collaborative Filtering

Collaborative Filtering ist eine weit verbreitete Technik in Empfehlungssystemen (Su und Khoshgoftaar 2009). Sie empfiehlt Artikel auf Grundlage der Vorlieben ähnlicher Benutzer*innen. Durch die Analyse von Verhaltensmustern und Ähnlichkeiten in den Interaktionen zwischen Benutzer*innen und Artikeln bzw. in den entsprechenden Bewertungen schlagen Collaborative Filtering Ansätze Artikel vor, die Benutzer*innen mit ähnlichen Vorlieben gefallen haben. Collaborative Filtering kann in zwei Hauptansätze unterteilt werden: nutzungsbezogenes und artikelbezogenes Collaborative Filtering (Aggarwal et al. 2016).

Nutzungsbezogenes Collaborative Filtering vergleicht die Vorlieben einer Zielnutzer*in mit denen ähnlicher Nutzer*innen, um Empfehlungen auszusprechen. Bei artikelbasiertem Collaborative Filtering hingegen werden ähnliche Artikel identifiziert und solche empfohlen, die denen ähneln, die der Zielbenutzer*in gefallen haben bzw. mit denen die Person zuvor interagiert hat.

Die Collaborative Filtering Ansätze haben sich zwar bei der Erstellung von personalisierten Empfehlungen als effektiv erwiesen, weisen aber auch Schwächen auf, wie zum Beispiel dem Kaltstartproblem (Schein et al. 2002), d.h. Schwierigkeiten bei der Empfehlung neuer oder inaktiver Benutzer*innen, und der Gefahr der Bildung von Filterblasen (Nguyen et al. 2014), d.h. Einschränkung der Empfehlungen auf eine begrenzte Auswahl. Auch sind für jede Nutzer*in in der Regel nur für wenige Artikel die Präferenzen bekannt, sodass man sehr viele Nutzende benötigt, um passende Empfehlungen aussprechen zu können.

2.1.2 Content-based Filtering

Bei Content-based Filtering werden Elemente auf der Grundlage ihrer Attribute oder Merkmale empfohlen. Solche Systeme analysieren den Inhalt oder die Merkmale bewerteter Artikel und gleichen sie mit den Präferenzen der Benutzer*in hinsichtlich anderer Artikel mit ähnlichen Merkmalen ab (Pazzani 1999).

Ein Vorteil inhaltsbasierter Methoden ist, dass Empfehlungen für neue Artikel gegeben werden können, auch wenn für diese noch nicht ausreichend Nutzungsinteraktionen vorhanden sind. Indem bewertete Artikel mit ähnlichen Attributen zur Bestimmung der Präferenz der nutzenden Person herangezogen werden, kann Content-based Filtering dennoch relevante Empfehlungen vorschlagen (Aggarwal et al. 2016).

Schwächen des Ansatzes sind nach Aggarwal et al. (2016) etwa die Abhängigkeit von Artikelmerkmalen wie Schlüsselwörtern oder anderem Inhalt, die zu offensichtlichen Empfehlungen und damit zu einer Verringerung der Vielfalt der empfohlenen Artikel führen kann. Außerdem setzt der Ansatz voraus, auf vorherige Bewertungen bzw. Interaktionen der gleichen Nutzer*in zugreifen zu können.

2.1.3 Hybride Ansätze

Hybride Empfehlungssysteme zeichnen sich durch die Kombination von Collaborative und Content-based Filtering aus. Diese Systeme nutzen die Vorteile beider Methoden, um genauere und vielfältigere Empfehlungen zu liefern (Burke 2002). Durch die Kombination von Nutzungspräferenzen (Collaborative Filtering) und Benutzer- und Artikelmerkmalen (Content-based Filtering) zielen hybride Ansätze darauf ab, die Empfehlungsqualität weiter zu verbessern, indem sie z. B. fehlende Bewertungen für das Collaborative Filtering einzelner Nutzer*innen auf Basis des Inhaltes vorhersagen und so die Einschränkungen der einzelnen Techniken vermeiden (Balabanović und Shoham 1997).

2.2 Betrachtung ethischer Aspekte

In den letzten Jahren hat die wachsende Bedeutung datengesteuerter Technologien auf der Grundlage des maschinellen Lernens Bedenken hinsichtlich ihrer Fairness, Transparenz und möglichen Voreingenommenheit aufgeworfen (Mehrabi et al. 2021):

Maschinelles Lernen, wie es für Empfehlungssysteme, aber auch im breiteren Kontext der Künstlichen Intelligenz (KI) verwendet wird, dient dazu, in Daten Muster zu bestimmen oder aus Daten Modelle zu lernen und diese für Vorhersagen oder Empfehlungen zu nutzen. Wenn die Trainingsdaten jedoch verzerrt sind, werden diese Verzerrungen auch vom Modell übernommen und spiegeln sich in dessen Ergebnissen wider (Chen et al. 2023).

Dies hat dazu geführt, dass sich eine wachsende Forschungsgemeinschaft mit den ethischen Auswirkungen von Systemen des maschinellen Lernens be-

schäftigt. In diesem Zusammenhang bietet die ACM Conference on Fairness, Accountability, and Transparency (ACM FAccT) (Fox et al. 2023) als interdisziplinäre Konferenz eine Plattform für Wissenschaftler*innen aus verschiedenen Bereichen wie Informatik, Recht, Sozial- und Geisteswissenschaften, um sich mit den ethischen Herausforderungen der KI als sozio-technologisches System auseinanderzusetzen, und hat die in ihrem Titel formulierten Aspekte von Fairness, Verantwortlichkeit und Transparenz als mögliche Teilperspektiven verantwortungsvoller KI benannt.

Im Zusammenhang mit Empfehlungssystemen zielen Workshops und Tutorials wie FAccTRec (Fairness, Accountability, and Transparency in Recommender Systems) (FAccTRec 2022) oder Fairness and discrimination in recommendation and retrieval (Ekstrand, Burke und Diaz 2019) darauf ab, speziell diese Fragestellungen für Empfehlungssysteme zu erörtern, und folgen damit dem Vorbild von ACM FAccT, um die jeweiligen Aspekte speziell für den Anwendungsfall verantwortungsvoller Empfehlungssysteme zu untersuchen. Dabei werden die im Titel genannten Konzepte der Fairness, der Rechenschaftspflicht und der Transparenz um weitere wichtige Themen wie Verantwortung, Sicherheit, Compliance und die Auswirkungen fairnessbewusster und verantwortungsvoller Empfehlungen in Industrie und Forschung erweitert (FAccTRec 2022).

3. Responsible Recommender Design

In diesem Abschnitt stellen wir ein medizinisches Empfehlungssystem aus praktischer Sicht vor und erörtern die ethischen und rechtlichen Implikationen des allgemeinen Umfelds und der Designentscheidungen für das Empfehlungssystem.

3.1 Recommender Use Case

In Krankenhäusern gibt es mehrere Abteilungen, die auf bestimmte Diagnosen spezialisiert sind und anderen Abteilungen Dienstleistungen anbieten, beispielsweise Radiologie oder Labore. Um eine Diagnose für eine Patient*in zu erstellen, stellt die federführende Abteilung des Falles eine Anfrage an eine andere Abteilung, mit der sie eine bestimmte Untersuchung anfordert. Diese Untersuchung wird dann in der Abteilung durchgeführt, die die diagnostische Leistung anbietet, und die Ergebnisse werden an die anfordernde Abteilung

zurückgemeldet. Während in der Vergangenheit diese Diagnostikanfragen überwiegend telefonisch organisiert wurden, haben heute vor allem größere Krankenhäuser diesen Prozess mit IT-Lösungen digitalisiert. Am in diesem Beispiel betrachteten Klinikum können Ärzt*innen auf diese Weise Diagnosen aus der Radiologie anfordern. Für einen effizienten Anforderungsprozess hat sich das Klinikum entschieden, nur eine grobe Diagnosekategorie wie z.B. Röntgen des Handgelenks definieren zu lassen, anstatt das gesamte Spektrum der rund 2000 möglichen Untersuchungen darzustellen, die durchgeführt und abgerechnet werden können. Für die Dokumentation und Abrechnung muss jedoch das genaue diagnostische Verfahren erfasst werden. Deshalb beschäftigt das Klinikum medizinische Fachangestellte, die die Anfrage verfeinern, bevor sie diese an die entsprechende Abteilung weiterleiten. Dieser Ansatz hat zwei Nachteile, die durch Empfehlungssysteme adressiert werden können. Erstens nimmt die Verfeinerung durch das medizinische Personal Zeit in Anspruch, was bei dringenden Fällen im Gegensatz zu einem direkten Kontakt zwischen anfordernder und leistender Abteilung ein Problem darstellen kann, insbesondere wenn es Rückfragen an die anfordernde Abteilung gibt. Zweitens wird die Kapazität des Personals für eine Aufgabe gebunden, die für den Prozess nicht zwingend erforderlich ist, und die ansonsten für wertvollere Aufgaben wie die ärztliche Versorgung der Patient*innen verwendet werden könnte.

Die Einbindung von Empfehlungssystemen in diesen Prozess kann dazu beitragen, diesen Prozess effizienter zu gestalten. Dafür können sie in zwei Prozessschritten eingesetzt werden. Erstens als Hilfsmittel für das medizinische Fachpersonal, was den Prozess der Verfeinerung von Anfragen effizienter macht. Zweitens als Hilfsmittel für die anfordernden Ärzt*innen, was es ihnen ermöglicht, Diagnosen feingranularer und präziser anzufordern, ohne sie dabei mit einer Vielzahl möglicher Optionen und einem zeitaufwändigen Auswahlprozess zu belasten.

Neben diesen Prozessschritten sind auch andere technische Überlegungen für die Gestaltung des Empfehlungssystems und für die sich daraus ergebenden möglichen ethischen Überlegungen von Bedeutung. So stellt sich beispielsweise die Frage, auf welcher Datenbasis das Empfehlungssystem trainiert werden soll. Durchgeführte und abgerechnete Diagnosen werden hier seit langem erfasst, da sie aus buchhalterischen Gründen digital dokumentiert werden müssen. Sie stellen eine relativ saubere Datenbasis dar, da von der Beantragung bis zur Durchführung der Untersuchung mehrere Expert*innen beteiligt waren, mögliche Fehler zu korrigieren. Nachteilig ist, dass viele die-

ser Aufzeichnungen nicht mit den ursprünglichen Anträgen und den ihnen zugrunde liegenden Entscheidungen in Verbindung gebracht werden können, da sie möglicherweise in der Vergangenheit telefonisch gestellt wurden. Außerdem spiegeln sie nicht die von der anfragenden Stelle bevorzugte Granularität wider, da sie hauptsächlich der Abrechnung der Leistungen dienen. Auf der anderen Seite sind die auf der anfordernden Seite verfügbaren Daten meist unstrukturiert und auf recht grobe Diagnosekategorien beschränkt. Zudem ist die Qualität der Daten nicht gesichert, da beispielsweise relevante Anteile der Anfragen von unerfahrenen Ärzt*innen generiert werden können, die in nachfolgenden Prozessschritten von Expert*innen für die jeweiligen Diagnosen verfeinert werden können. Dies wird auch im Hinblick auf den Recommender-Ansatz relevant, da zum Beispiel Ansätze, die auf Collaborative Filtering basieren, unter der Datenqualität leiden könnten und dadurch möglicherweise nicht optimale Untersuchungen empfehlen. Dies erfordert wiederum eine nachträgliche Verfeinerung der Anfragen, wohingegen wissensbasierte Ansätze, die auf klinischen Leitlinien basieren, in ähnlichen Anwendungen nachweislich bessere Ergebnisse liefern (Mei et al. 2015).

Unter dem Gesichtspunkt der Akzeptanz bringt dies ebenfalls Herausforderungen mit sich, da sich die anfordernden Ärzt*innen durch Vorschläge, die über den ursprünglich beabsichtigten Rahmen hinausgehen, bevormundet fühlen könnten, während dies andererseits das Potenzial bietet, die Qualität der Anfragen zu verbessern und die Anzahl der erforderlichen Nachbesserungen zu verringern. Das Anbieten von Empfehlungen, die die Intention des Anfragenden zu genau modellieren, wirft dagegen Fragen der (gefühlten) Verantwortlichkeit auf, da die anfragenden Ärzt*innen in Verhaltensmuster verfallen könnten, die dazu führen, sich zu sehr auf die Empfehlung zu verlassen.

Neben diesen Überlegungen wirft die Entwicklung und der Einsatz eines Empfehlungssystems in diesem sensiblen medizinischen Kontext auch mehrere Fragen in Hinblick auf Fairness, Verantwortlichkeit und Transparenz auf, die in den folgenden Abschnitten behandelt werden.

3.2 Verantwortlichkeit

Die im vorangegangenen Abschnitt skizzierten Aspekte der Verantwortlichkeit für ein Empfehlungssystem können aus zwei Perspektiven mit auf den ersten Blick widersprüchlichen Sichtweisen betrachtet werden. Aus praktischer Sicht entlastet ein Empfehlungssystem, das die Grundlage für die Beantragung der gewünschten Untersuchung bietet, das Assistenzpersonal mit dem Potenzi-

al, es in der Prozesskette letztlich einzusparen. Aus sozio-ökonomischer Sicht geht es um die Verantwortung für diese Mitarbeiter*innen, da sich für das Krankenhaus Möglichkeiten ergeben könnten, Personal abzubauen, um Kosten zu sparen. Diese schwerwiegenden Folgen für die Betroffenen könnten gar den Einsatz als »verantwortungsvolles« Empfehlungssystem nicht mehr rechtfertigen.

Aus einer anderen Perspektive kann die gleiche Situation zu einem anderen Ergebnis führen. Wenn das Empfehlungssystem es ermöglicht, eine große Anzahl von Standardfällen direkt anzufordern, kommt die Zeitersparnis den Patient*innen zugute. Das Assistenzpersonal kann sich dann auf schwierige Fälle konzentrieren, Anfragen bearbeiten, die einer weiteren Klärung bedürfen, oder mit anderen Aufgaben betraut werden, die die Versorgung aus Sicht der Patient*innen insgesamt verbessern.

Diese Überlegungen deuten darauf hin, dass das tatsächliche Ergebnis hinsichtlich der Verantwortungsaspekte von den Entscheidungen der Krankenhausleitung abhängt, was sich mit der kürzlich von Gansky und McDonald (2022) geführten Diskussion deckt, die anmerken, dass der organisatorische Kontext, in dem das System eingesetzt werden soll, berücksichtigt werden muss.

Daher scheint der Einsatz eines Empfehlungssystems zum Wohle der Patient*innen, die in einer Krankenhausumgebung im Mittelpunkt stehen, vertretbar, wenn Compliance- und Sicherheitsaspekte eingehalten werden.

3.3 Fairness

Fairness von Empfehlungssystemen kann aus einer prozeduralen Perspektive (Lee et al. 2019) mit einem Fokus auf Fairness innerhalb des Entscheidungsprozesses oder aus einer ergebnisorientierten Perspektive der Behandlung ähnlicher Individuen oder Gruppen (Biega, Gummadi und Weikum 2018) bewertet werden. Während Fairness im Allgemeinen mit der Verringerung oder Vermeidung von Verzerrungen im Entscheidungsprozess des maschinellen Lernens verbunden ist, die oft durch die Daten eingeführt werden (Zou und Schiebinger 2018), kann sie aus einer prozeduralen Perspektive (Lee et al. 2019) mit einem Fokus auf Fairness innerhalb des Entscheidungsprozesses oder aus einer ergebnisorientierten Perspektive bei der Behandlung ähnlicher Individuen oder Gruppen (Biega, Gummadi und Weikum 2018) angegangen werden.

Der Unterschied zwischen einer prozeduralen und einer ergebnisorientierten Perspektive kann zu unterschiedlichen Bewertungen in Bezug auf ethi-

sche Belange und somit zu unterschiedlichen Ergebnissen führen. Tatsächlich kann verfahrensbezogene Fairness zu Ergebnissen führen, die als ungerecht empfunden werden, und umgekehrt (Lee et al. 2019). Speziell im Bereich der Gesundheitsversorgung wurde eine ähnliche Frage von Tsuchiya und Dolan (2009) untersucht, die die Sichtweise der Öffentlichkeit auf ergebnis- und gewinnbezogene Fairness betrachtet haben. Sie warfen Fragen bezüglich der Verzerrung der wahrgenommenen Fairness in der Gesellschaft auf. In ihrer Studie zeigen Tsuchiya und Dolan (2009), dass die Mehrheit der Befragten in Bezug auf den sozialen Status eine ergebnisorientierte Fairness im medizinischen Kontext bevorzugt, während eine beträchtliche Minderheit eine gewinnorientierte Fairness bevorzugt. Für die ergebnisorientierte Perspektive ist jedoch die Definition des Ergebnisses von großer Bedeutung. Offensichtlich kann die Empfehlung einer Untersuchung nicht als Ergebnis angesehen werden, da unterschiedliche (potenziell geschützte) Untergruppen, beispielsweise Kinder und ältere Menschen, aus medizinischer Sicht unterschiedliche Untersuchungen benötigen. Die Definition des Ergebnisses ist auch für die Bewertung der Empfehlung selbst von Bedeutung, da das Lernen aus in der Vergangenheit durchgeführten Untersuchungen als Grundwahrheit nicht unbedingt die beste Wahl für die Patient*in widerspiegelt und den Erfolg nachfolgender Behandlungen nicht bewerten kann. Eine Lösung, wie sie von Mei et al. (2015) vorgeschlagen wurde, ist die Verwendung einer kohortenstudienbasierten Auswertung zur Bewertung des Erfolgs der empfohlenen Behandlungen und zur Bewertung der Fairness. In der hier betrachteten Fragestellung ist das Ergebnis jedoch schwer zu definieren, da selbst einfache Kriterien, wie die Wartezeit bis zur Diagnosestellung, durch gruppenspezifische medizinische Aspekte, wie beispielsweise die Dringlichkeit, verzerrt werden können und sich diese Verzerrung auch in den Daten wiederfindet. Zusätzlich zu solchen gerechtfertigten und gewünschten Verzerrungen können sich in den Daten aber auch implizite Verzerrungen wiederfinden, die im Sinne eines fairen Entscheidungssystems nicht durch das System übernommen werden sollten.

Dies zeigt, dass bestimmte Kompromisse in Bezug auf Fairness und Ethik aus fachspezifischer Sicht notwendig sind, um die medizinische und ökonomische Realität widerzuspiegeln, wie sie in der täglichen medizinischen Praxis benötigt wird (Beauchamp und Childress 1994), und um damit die Akzeptanz für alle Beteiligten zu gewährleisten (Milano, Taddeo und Floridi 2020). Dennoch muss die Prüfung dieser Kompromisse auf ihre klinische und ethische Validität in den Bewertungsprozess des Empfehlungssystems einbezogen werden.

Aus algorithmischer Sicht spielt die Unverzerrtheit der Daten in Hinblick auf Ansätze, die Collaborative Filtering verwenden, eine größere Rolle als bei Content-basierten Verfahren, insbesondere solchen, die auf medizinischem Wissen basieren, da sich potentielle Fairnessprobleme aus der Vergangenheit, die in den Daten vorkommen, durch den inhärenten Zusammenhang mit ähnlichen Fällen wiederholen würden. Zahlreiche Forschungsergebnisse versuchen außerdem auf algorithmischer Ebene unterschiedliche Perspektiven von Fairness sicherzustellen (Wang et al. 2023), die für die Umsetzung eines medizinischen Empfehlungssystems Beachtung finden sollten.

3.4 Rechenschaftspflicht

Die Rechenschaftspflicht im Zusammenhang mit Systemen zur Empfehlung von Untersuchungen konzentriert sich auf die Frage, wer für den potenziellen Schaden, den das komplexe sozio-technische System einschließlich der anfordernden Ärzt*innen, des Empfehlungssystems und seiner Entwickler*innen, des medizinischen Assistenzpersonals und der ausführenden Ärzt*innen unter Umständen verursacht, verantwortlich ist und wer sich dafür verantwortlich fühlt. In einer aktuellen Studie des Europäischen Parlamentarischen Forschungsdienstes über KI im Gesundheitswesen (Lekadir et al. 2022) wird der Bedarf an neuen Mechanismen und Rahmenbedingungen zur Gewährleistung einer angemessenen Rechenschaftspflicht bei medizinischer KI festgestellt. Lücken in der KI-Rechenschaftspflicht finden sich in drei Aspekten: Aus rechtlicher Sicht fehlende Regelungen und Definitionen für die Rechenschaftspflicht, die Schwierigkeit, Rollen und damit verbundene Verantwortlichkeiten innerhalb des soziotechnischen Prozesses zu definieren, und das Fehlen einer rechtlichen und ethischen Governance für KI-Entwicklende und -Herstellende. Während die meisten vorgeschlagenen Maßnahmen eine regulatorische Sichtweise einnehmen, schlägt eine der Maßnahmen vor, Prozesse zu implementieren, um die Rollen von KI und Nutzende zu identifizieren, wenn KI-basierte Entscheidungen Patient*innen schaden könnten und die Verantwortlichkeiten explizit zu machen. In ähnlicher Weise schlagen Habli, Lawton und Porter (2020) vor, KI-Entwicklende und -Ingenieur*innen einzubeziehen, wenn es darum geht, die Verantwortlichkeit für Entscheidungen im Rahmen des KI-gestützten Prozesses zu bewerten. Darüber hinaus erörtern sie, dass die Rechenschaftspflicht unmittelbar mit Sicherheit verbunden ist, die während der KI-Entwicklung nicht vollständig vorhersehbar

ist, insbesondere im Hinblick auf mögliche Verhaltensanpassungen von Ärzt*innen, Patient*innen und dem System selbst.

Da die Ärzt*innen die endgültige Entscheidung über die Annahme oder Ablehnung von Empfehlungen treffen, schlagen wir vor, das Bewusstsein für Verantwortlichkeit und Sicherheitsbedenken als Teil des Anforderungsprozesses zu schärfen, indem die Annahmequote der vorgeschlagenen Empfehlung beobachtet wird. Wenn diese Auswertung darauf hindeutet, dass sich die anfragende Person zu sehr auf die Empfehlung verlässt, können Schritte zur Sensibilisierung eingeleitet werden, beispielsweise durch die Empfehlung einer unzulässigen Diagnose, die, wenn sie unreflektiert akzeptiert wird, mit einer Warnmeldung gestoppt wird.

3.5 Transparenz

Transparenz wird von Smith (2021) als Schlüsselaspekt für die Rechenschaftspflicht und damit für die Akzeptanz identifiziert, die argumentiert, dass Ärzt*innen für den Einsatz in der klinischen Praxis über ihre Entscheidung Rechenschaft ablegen müssen und intransparente KI-Systeme ablehnen werden, da sie deren Ergebnis nicht nachvollziehen können. Andererseits behaupten Clement, Ren und Curley (2021) in ihrer empirischen Studie, dass Transparenz die Akzeptanz minderwertiger Empfehlungen begünstigt, also eine Verhaltensänderung bewirkt, indem sie potentiell falsche Modellempfehlungen mit plausibel klingenden Erklärungen versieht.

Da die von den Ärzt*innen formulierten Anforderungen auf medizinischen Entscheidungen beruhen, dient das Empfehlungssystem in unserem Szenario in erster Linie dazu, den Auswahlprozess und nicht den Entscheidungsprozess zu verbessern. Daher müssen die Ärzt*innen in diesem Rahmen nicht für das Ergebnis des Systems, also die Empfehlung, die Verantwortung übernehmen, sondern für ihre Entscheidung. Eine von der eigentlichen Absicht abweichende Empfehlung muss einer medizinischen Bewertung unterzogen werden, bevor sie als praktikable Option eingestuft wird. Eine automatisierte Rationalisierung oder Erklärung der Empfehlung könnte daher den Bewertungsprozess abkürzen und damit die notwendige Sorgfalt untergraben, was auch mit der allgemeineren Beobachtung der Aufwertung algorithmischer Entscheidungen in Zusammenhang steht (Logg, Minson und Moore 2019). Eine weniger eingreifende Form der Transparenz, die Ärzt*innen die notwendigen Informationen liefert, ohne die eigentliche Argumentation vorwegzunehmen, kann daher besser geeignet sein. Die Bereitstellung

ähnlicher Fälle zum Vergleich für einen auf Collaborative Filtering basierenden Ansatz oder die Unterstützung wissensbasierter Empfehlungen durch klinische Leitlinien könnten ein vielversprechender Kompromiss zwischen der Undurchsichtigkeit der Empfehlungen und feinkörnigen Begründungen darstellen, ohne zu unerwünschten Verhaltensänderungen zu führen.

3.6 Compliance

Was die Einhaltung der Vorschriften betrifft, so gelten in Deutschland und der Europäischen Union für medizinische KI die 2017/746 In Vitro Diagnostic Medical Devices Regulation (IVDR) und die 2017/745 Medical Devices Regulations (MDR) (Lekadir et al. 2022), wobei letztere eher für die Festlegung von Diagnoseempfehlungen gilt. Kiseleva (2020) kommt zu dem Schluss, dass diese Verordnung als anfänglicher rechtlicher Rahmen dienen kann, jedoch in Bezug auf Transparenz und Rechenschaftspflicht erweitert werden muss. Dies wird durch die Notwendigkeit einer Risikobewertung im Vorschlag des Forschungsdienstes des Europäischen Parlaments über Künstliche Intelligenz im Gesundheitswesen (Lekadir et al. 2022) ergänzt, der in Übereinstimmung mit dem Vorschlag der Europäischen Kommission für eine KI-Regulierung zusätzlich vorschlägt, Anwendungen nach Risikostufen von inakzeptabel bis minimal einzustufen. Für Empfehlungssysteme wurde die Anwendbarkeit dieser Regelung kürzlich von Schwemer (2021) diskutiert, wobei die Anwendung im medizinischen Kontext als potenzielle Hochrisikoanwendungen eingestuft werden dürften (Lekadir et al. 2022). Aus praktischer Sicht bietet die Initiative FUTURE-AI (Lekadir et al. 2021) Leitlinien und Best Practices für vertrauenswürdige KI in der Medizin, die bei der Gestaltung eines Empfehlungssystems berücksichtigt werden sollten.

3.7 Sicherheit

Der Aspekt der Sicherheit ist nicht nur mit der eigentlichen Performance des Empfehlungssystems verbunden, was die offensichtliche Fragestellung darstellt, wie geeignet die Empfehlungen sind und ob der Einsatz des Systems positive Auswirkungen hat, oder ob falsche oder ungenaue Empfehlungen eines schwachen Empfehlungssystems einen negativen Einfluss haben können, etwa durch zusätzlichen Zeitaufwand bei der Verfeinerung der Anfrage bis hin zu nicht optimalen diagnostischen Entscheidungen, die wichtige Gesundheitszustände übersehen und somit direkt oder indirekt Schaden

für die Patient*innen verursachen (Lekadir et al. 2022). Stattdessen kann die Sicherheit auch aus einer nutzungszentrierten Perspektive angegangen werden, wie im Zusammenhang der Rechenschaftspflicht erörtert, wo der Missbrauch eines Systems zu Sicherheitsproblemen führen kann, sowie aus einer datenorientierten Perspektive, beispielsweise, wenn die Daten falsche Informationen oder Rauschen enthalten oder die Datenpunkte aus einer auf klinischen Leitlinien basierenden Perspektive veraltet sind. Dies spiegelt sich auch in praktischen Richtlinien und Vorschriften wider, da Sicherheit am besten ganzheitlich betrachtet wird, von der Datenerfassung, der Annotation, über das Systemdesign und der Evaluierung bis hin zum sozio-technischen und organisatorischen Kontext, in dem es verwendet wird (Lekadir et al. 2021; Dobbe 2022) und in diesem Komplex entsprechend geprüft wird (Falco et al. 2021).

Für die Bewertung der Sicherheit unseres Empfehlungssystems bedeutet dies, dass wir uns nicht nur auf Leistungskennzahlen verlassen können, zumal die Daten selbst anfällig für Fehler und Rauschen sein können. Während die Leistung des Systems in einem Bereich liegen muss, in dem der potenzielle Nutzen die Sicherheitsrisiken überwiegt, um brauchbar zu sein, muss der Nutzen und die Sicherheit aus einer Ergebnisperspektive konsequent überwacht werden, insbesondere während des Betriebs.

3.8 Zusammenfassung und Diskussion

Unter Berücksichtigung der vorangegangenen Überlegungen, die wir im Folgenden nochmals zusammenfassen, könnte der Einsatz eines verantwortungsvollen Empfehlungssystems zum Wohle der Patient*innen vertretbar sein, wenn Compliance- und Sicherheitsaspekte eingehalten werden können. Hinsichtlich Fairness scheint das Ergebnis in der hier betrachteten Fragestellung schwer bewertbar zu sein, da selbst einfache Kriterien wie die Wartezeit bis zur Diagnosestellung durch gruppenspezifische medizinische Aspekte, wie beispielsweise die Dringlichkeit, verzerrt werden können. Dies zeigt, dass bestimmte Kompromisse in Bezug auf Fairness und Ethik aus fachspezifischer Sicht notwendig sind, um die medizinische und ökonomische Realität widerzuspiegeln (Beauchamp und Childress 1994) und damit die Akzeptanz für alle Beteiligten zu gewährleisten (Milano, Taddeo und Floridi 2020). Daher muss die Prüfung dieser Kompromisse auf ihre klinische und ethische Validität in den Bewertungsprozess des Empfehlungssystems einbezogen werden. Die Algorithmik des Empfehlungssystems ist nicht in der Lage, diese

Balance automatisch zu gewährleisten und bedarf daher Anpassungen, falls die Fairness nicht sichergestellt ist.

Da die Ärzt*innen die endgültige Entscheidung über die Annahme oder Ablehnung von Empfehlungen treffen, sind sie auch verantwortlich und damit rechenschaftspflichtig. Wir schlagen deshalb vor, das Bewusstsein für Verantwortlichkeit und Sicherheitsbedenken als Teil des Anforderungsprozesses immer wieder zu schärfen, indem beispielsweise die Annahmequote der vorgeschlagenen Empfehlung beobachtet und hinterfragt wird. Hinsichtlich der Transparenz könnten plausibel klingende Erklärungen die nötige Sorgfalt, diese nicht unhinterfragt zu übernehmen, einschränken. Eine weniger eingreifende Form der Transparenz, die Ärzt*innen die notwendigen Informationen liefert, ohne die eigentliche Argumentation vorwegzunehmen, dürfte daher besser geeignet sein, um trotzdem Transparenzanforderungen an das Systems zu adressieren. Nach dem Vorschlag des Forschungsdienstes des Europäischen Parlaments über Künstliche Intelligenz im Gesundheitswesen (Lekadir et al. 2022) dürfte die Anwendung von Empfehlungssystemen im medizinischen Kontext als potenzielle Hochrisikoanwendungen eingestuft werden (Lekadir et al. 2022). Aus praktischer Sicht bietet die Initiative FUTURE-AI (Lekadir et al. 2021) Leitlinien und Best Practices für vertrauenswürdige KI in der Medizin, die bei der Gestaltung eines Empfehlungssystems berücksichtigt werden sollten, um den hohen Anforderungen als Hochrisikoanwendung gerecht zu werden. Sicherheitsaspekte sollten am besten ganzheitlich betrachtet werden, von der Datenerfassung, der Annotation, über das Systemdesign und der Evaluierung bis hin zum sozio-technischen und organisatorischen Kontext, in dem es verwendet wird (Lekadir et al. 2021; Dobbe 2022) und in diesem Komplex entsprechend geprüft werden (Falco et al. 2021), insbesondere auch während des Betriebs.

4. Fazit

In dieser Arbeit haben wir ein Empfehlungssystem für medizinische Untersuchungen in einem anwendungsorientierten Kontext skizziert. Wir haben die Implikationen der Designentscheidungen, des Anwendungsfalls und des Systems im Hinblick auf Verantwortung, Fairness, Rechenschaftspflicht, Transparenz, Compliance und Sicherheit erörtert und einen Überblick über die Möglichkeiten gegeben, wie diese Implikationen und Probleme aus praktischer Sicht adressiert werden können. Im Kontext unseres Empfeh-

lungssystems konnten wir damit wichtige ethische Aspekte, die für eine verantwortungsvolle Gestaltung und Implementierung relevant sind, operationalisieren und im Kontext bisheriger Arbeiten bewerten, und hoffen, damit Impulse für die weitere Entwicklung eines solchen Systems und deren Umsetzung in der Praxis zu geben.

Danksagung

Diese Forschung wurde durch das Bayerische Staatsministerium für Wissenschaft und Kunst im Projekt »Digitalisierungszentrum für Präzisions und Telemedizin« (DZ.PTM) im Rahmen des Masterplans »BAYERN DIGITAL II« gefördert.

Literatur

Adomavicius, Gediminas, and Alexander Tuzhilin. 2005. »Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions«. *IEEE transactions on knowledge and data engineering* 17 (6): 734–749.

Aggarwal, Charu C., et al. 2016. *Recommender systems*. Bd. 1. Springer.

Balabanović, Marko, and Yoav Shoham. 1997. »Fab: content-based, collaborative recommendation«. *Communications of the ACM* 40 (3): 66–72.

Beauchamp, Tom L. and James F. Childress 1994. Principles of biomedical ethics. Oxford University Press.

Biega, Asia J., Krishna P. Gummadi and Gerhard Weikum. 2018. »Equity of attention: Amortizing individual fairness in rankings«. In *The 41st international acm sigir conference on research & development in information retrieval*, 405–414.

Bobadilla, Jesús, Fernando Ortega, Antonio Hernando and Abraham Gutiérrez. 2013. »Recommender systems survey«. *Knowledge-based systems* 46: 109–132.

Burke, Robin. 2002. »Hybrid recommender systems: Survey and experiments«. In *User Modeling and User-Adapted Interaction* (12): 331–370.

Chen, Jiawei, Hande Dong, Xiang Wang, Fuli Feng, Meng Wang and Xiangnan He. 2023. »Bias and debias in recommender system: A survey and future directions«. *ACM Transactions on Information Systems* 41 (3): 1–39.

Clement, Jeffrey, Yuqing Ching Ren and Shawn Curley. 2021. »Increasing System Transparency About Medical AI Recommendations May Not Improve Clinical Experts' Decision Quality«. Available at SSRN 3961156.

Dallmann, Alexander, Johannes Kohlmann, Daniel Zoller and Andreas Hotho. 2021. »Sequential Item Recommendation in the MOBA Game Dota 2«. In *2021 International Conference on Data Mining Workshops (ICDMW)*, 10–17. IEEE.

Dobbe, Roel. 2022. »System Safety and Artificial Intelligence«. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, 1584–1584.

Ekstrand, Michael D., Robin Burke and Fernando Diaz. 2019. »Fairness and discrimination in recommendation and retrieval«. In *Proceedings of the 13th ACM Conference on Recommender Systems*, 576–577.

FAccTRec. 2022. *RecSys – ACM Recommender Systems*. Accessed February 20, 2023. https://recsys.acm.org/recsys22/facctrec/.

Falco, Gregory, Ben Shneiderman, Julia Badger, Ryan Carrier, Anton Dahbura, David Danks, Martin Eling, Alwyn Goodloe, Jerry Gupta, Christopher Hart et al. 2021. »Governing AI safety through independent audits«. *Nature Machine Intelligence* 3 (7): 566–571.

Fischer, Elisabeth, Daniel Zoller and Andreas Hotho. 2021. »Comparison of Transformer-Based Sequential ProductRecommendation Models for the Coveo Data Challenge«. *SIGIR Workshop On eCommerce* (July).

Fox, Sarah, Christina Harrington, Aziz Huq and Chenhao Tan, (Hg.). 2023. *FAccT '23: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. Chicago, IL, USA: Association for Computing Machinery.

Gansky, Ben, and Sean McDonald. 2022. »CounterFAccTual: How FAccT Undermines Its Organizing Principles«. *In 2022 ACM Conference on Fairness, Accountability, and Transparency*, 1982–1992. FAccT '22. Seoul, Republic of Korea: Association for Computing Machinery. DOI: https://doi.org/10.1145/3531146.3533241.

Gomez-Uribe, Carlos A., and Neil Hunt. 2015. »The netflix recommender system: Algorithms, business value, and innovation«. *ACM Transactions on Management Information Systems (TMIS)* 6 (4): 1–19.

Habli, Ibrahim, Tom Lawton and Zoe Porter. 2020. »Artificial intelligence in health care: accountability and safety«. *Bulletin of the World Health Organization* 98 (4): 251.

Jäschke, Robert, Leandro Balby Marinho, Andreas Hotho, Lars Schmidt-Thieme and Gerd Stumme. 2007. »Tag Recommendations in Folksonomies«. In *Knowledge Discovery in Databases: PKDD 2007. 11th European*

Conference on Principles and Practice of Knowledge Discovery in Databases, Warsaw, Poland, September 17–21, 2007, Proceedings, edited by Joost N. Kok, Jacek Koronacki, Ramon López de Mántaras, Stan Matwin, Dunja Mladenic and Andrzej Skowron, 4702: 506–514. Lecture Notes in Computer Science. Springer. https://www.kde.cs.uni-kassel.de/wp-content/uploads/hotho/pub/2007/kdml_recommender_final.pdf.

Kiseleva, Anastasiya. 2020. »AI as a Medical Device: Is It Enough to Ensure Performance Transparency and Accountability in Healthcare?« *European Pharmaceutical Law Review*, Nr. 1.

Lee, Min Kyung, Anuraag Jain, Hea Jin Cha, Shashank Ojha and Daniel Kusbit. 2019. »Procedural justice in algorithmic fairness: Leveraging transparency and outcome control for fair algorithmic mediation«. *Proceedings of the ACM on Human-Computer Interaction* 3 (CSCW): 1–26.

Lekadir, Karim, Richard Osuala, Catherine Gallin, Noussair Lazrak, Kaisar Kushibar, Gianna Tsakou, Susanna Aussó, Leonor Cerdá Alberich, Kostas Marias, Manolis Tsiknakis et al. 2021. »FUTURE-AI: Guiding Principles and Consensus Recommendations for Trustworthy Artificial Intelligence in Medical Imaging«. *arXiv preprint arXiv*: 2109.09658.

Lekadir, Karim, Gianluca Quaglio, Anna Tselioudis Garmendia and Catherine Gallin. 2022. *Principles of biomedical ethics*. European Parliamentary Research Service.

Logg, Jennifer M, Julia A. Minson and Don A Moore. 2019. »Algorithm appreciation: People prefer algorithmic to human judgment«. *Organizational Behavior and Human Decision Processes* 151: 90–103.

Mehrabi, Ninareh, Fred Morstatter, Nripsuta Saxena, Kristina Lerman and Aram Galstyan. 2021. »A survey on bias and fairness in machine learning«. *ACM Computing Surveys (CSUR)* 54 (6): 1–35.

Mei, Jing, Haifeng Liu, Xiang Li, Yiqin Yu and Guotong Xie. 2015. »Outcomedriven Evaluation Metrics for Treatment Recommendation Systems.« In *MIE*, 190–194.

Milano, Silvia, Mariarosaria Taddeo and Luciano Floridi. 2020. »Recommender systems and their ethical challenges«. *Ai & Society* 35 (4): 957–967.

Mooney, Raymond J., and Loriene Roy. 2000. »Content-based book recommending using learning for text categorization«. In *Proceedings of the fifth ACM conference on Digital libraries*, 195–204.

Nguyen, Tien T., Pik-Mai Hui, F. Maxwell Harper, Loren Terveen and Joseph A. Konstan. 2014. »Exploring the filter bubble: the effect of using recom-

mender systems on content diversity«. In *Proceedings of the 23rd international conference on World wide web*, 677–686.

Pazzani, Michael J. 1999. »A framework for collaborative, content-based and demographic filtering«. *Artificial intelligence review* 13: 393–408.

Pazzani, Michael J, and Daniel Billsus. 2007. »Content-based recommendation systems«. *The adaptive web: methods and strategies of web personalization*, 325–341.

Schein, Andrew I., Alexandrin Popescul, Lyle H. Ungar and David M. Pennock. 2002. »Methods and metrics for cold-start recommendations«. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, 253–260.

Schwemer, Sebastian Felix. 2021. »Recommender Systems in the EU: from Responsibility to Regulation?« In *FAccTRec Workshop*, Bd. 21.

Shi, Yue, Martha Larson and Alan Hanjalic. 2014. »Collaborative filtering beyond the user-item matrix: A survey of the state of the art and future challenges«. *ACM Computing Surveys (CSUR)* 47 (1): 1–45.

Smith, Helen. 2021. »Clinical AI: opacity, accountability, responsibility and liability«. *AI & SOCIETY* 36 (2): 535–545.

Song, Yading, Simon Dixon and Marcus Pearce. 2012. »A survey of music recommendation systems and future perspectives«. In *9th international symposium on computer music modeling and retrieval*, 4: 395–410.

Su, Xiaoyuan, and Taghi M. Khoshgoftaar. 2009. »A survey of collaborative filtering techniques«. *Advances in artificial intelligence* 2009.

Sun, Leilei, Chuanren Liu, Chonghui Guo, Hui Xiong and Yanming Xie. 2016. »Data-driven automatic treatment regimen development and recommendation«. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 1865–1874.

Tsuchiya, Aki, and Paul Dolan. 2009. »Equality of what in health? Distinguishing between outcome egalitarianism and gain egalitarianism«. *Health economics* 18 (2): 147–159.

Wang, Yifan, Weizhi Ma, Min Zhang, Yiqun Liu and Shaoping Ma. 2023. »A Survey on the Fairness of Recommender Systems«. *ACM Transactions on Information Systems* 41, Nr. 3 (July): 1–43. DOI: https://doi.org/10.1145/3547333.

Zhang, Shuai, Lina Yao, Aixin Sun and Yi Tay. 2019. »Deep learning based recommender system: A survey and new perspectives«. *ACM computing surveys (CSUR)* 52 (1): 1–38.

Zou, James, and Londa Schiebinger. 2018. »AI can be sexist and racist—it's time to make it fair.« *Nature* 559 (7714): 324–326.

Zu viel Gewissheit? Herausforderungen künstlich-intelligenter Gesundheitsprädiktionen für die öffentliche Gesundheitsversorgung

Ulrich Freiherr von Ulmenstein, Max Tretter, Christina Lauppert von Peharnik, David Ehrlich

Einleitung

Künstliche Intelligenz (KI) dringt kontinuierlich in immer mehr Lebensbereiche vor und löst dort umfassende Transformationen aus. Gerade auf epistemischer Ebene bergen KI-Systeme ein enormes Potenzial, signifikante Veränderungen zu bewirken (Schmidt 2022). Die ersten Vorboten dieser Veränderungen werden schon sichtbar in der Art und Weise, wie große Sprachmodelle und intelligente Suchmaschinen unseren Umgang mit Informationen beeinflussen und dabei unsere Vorstellungen von Wissen selbst neu prägen (Dabrock 2023). Doch die Potenziale der KI beschränken sich nicht auf den Alltagsgebrauch. Im Gesundheitswesen beispielsweise werden fortschrittliche KI-Anwendungen zur Analyse von Daten genutzt, um die Gesundheitsentwicklung von Einzelpersonen zu prognostizieren (Chaari 2019; Topol 2020). Im politischen Kontext werden KI-Systeme eingesetzt, um die möglichen Auswirkungen verschiedener Entscheidungen auf die Gesellschaft abzuschätzen (Valle-Cruz und Sandoval-Almazan 2018; Valle-Cruz et al. 2019).

Es ist unbestritten, dass KI-gestützte Prognosen zum aktuellen Zeitpunkt nicht vollständig zuverlässig sind. Die Möglichkeit, dass KI-Systeme Fehleinschätzungen treffen und prognostizierte Ereignisse nicht oder anders als erwartet eintreten, ist stets präsent. Im Endeffekt sind die Prognosen von KI-Systemen immer nur bedingt verlässlich. Doch auch die Gewissheiten, die sich aus solchen KI-Prädiktionen gewinnen lassen, können sich als äußerst hilfreich erweisen. Selbst wenn eine medizinische KI lediglich eine grobe Tendenz abgeben kann, wie sich die Gesundheit einer Person in den kommenden

Jahren voraussichtlich entwickeln wird, sind diese Vorhersagen oftmals nützlicher als ein komplettes Informationsvakuum – und trotz der Gefahr einer Fehleinschätzung – häufig verlässlicher als menschliche Schätzungen. So sehr KI-basierte Prognosen einerseits mit Vorsicht zu genießen und als grobe Annäherungen zu betrachten sind, deren Eintritt zwar nicht unwahrscheinlich ist, die aber dennoch nicht als absolut verlässlich eingestuft werden sollten, so wertvoll können sich derlei KI-getriebene Gewissheitsgewinne andererseits doch erweisen (Tretter 2023).

In diesem Beitrag werden wir die Überlegungen hinsichtlich des Zusammenhangs zwischen KI und Gewissheit weiterführen. Anhand eines gedanklichen Experiments wollen wir untersuchen, welche potenziellen Herausforderungen auf das öffentliche Gesundheitssystem zukommen könnten, sollten KI-Anwendungen in der Lage sein, die Gesundheitsentwicklung von Einzelpersonen mit außerordentlicher Präzision vorherzusagen. Dabei gehen wir von einer Vorhersagegenauigkeit aus, die so hoch ist, dass eventuelle Fehlprognosen als vernachlässigbar gelten können.

Obwohl KI-basierte Gesundheitsprädiktionen auf der einen Seite enorme Chancen bergen, besteht auf der anderen Seite die Befürchtung, dass der Einsatz von KI zur Gewissheitsschaffung in Gesundheitskontexten bestehende Finanzierungsprobleme in dualen öffentlich-privaten Gesundheitssystemen verschärfen könnte (Corea 2019). Dieser Aspekt wird im Folgenden tiefergehend behandelt, indem wir aufzeigen, wie der Einsatz von prädiktiver und gewissheitsschaffender KI im Gesundheitswesen adverse Selektionsdynamiken verstärken kann.

Um diese These zu untermauern, klären wir zunächst relevante konzeptionelle Begriffe und erläutern dann, welche epistemischen Verschiebungen von Ungewissheit und Gewissheit der Einsatz von KI im Gesundheitsbereich bereits jetzt und umso mehr in der Zukunft befördern kann. Am Beispiel des deutschen Gesundheitssystems analysieren wir dann die komplexen Auswirkungen neugewonnener Gesundheitsgewissheiten unter den systemischen Bedingungen des deutschen Gesundheitssystems (insbesondere den Wahlmöglichkeiten der Versicherungsformen) auf die Zusammensetzung der Versichertengemeinschaften. Dabei wird deutlich, dass Personen mit positiven Gesundheitsprognosen einen starken Anreiz haben könnten, von der gesetzlichen Krankenversicherung (GKV) in die private Krankenversicherung (PKV) zu wechseln. Eine solche Dynamik könnte die Finanzierung der GKV bedrohen. Nach einer Zusammenfassung der bisherigen Erkenntnisse geben wir einen kurzen Ausblick darauf, wie man diesem Problem begegnen könnte.

Konzeptions- und Begriffsklärungen

Den anzustellenden Untersuchungen seien einige begriffskonzeptionelle Erläuterungen zu *Gewissheit* und *Ungewissheit* sowie *KI* und *Gesundheitssystem* vorangestellt. Sie dienen der Bereitstellung eines einheitlichen Betrachtungsrahmens.

Wir legen dem Beitrag ein pragmatisches Verständnis von *Ungewissheit* und *Gewissheit* zu Grunde. Dieses wird am besten deutlich, wenn man es vom Zentralkonzept des Pragmatismus, dem Handlungsbegriff, her versteht. Um das eigene Handeln vernünftig planen und informierte Handlungsentscheidungen treffen zu können, brauchen Personen erstens eine Kenntnis ihrer gegenwärtigen Situation. Das bedeutet, sie brauchen eine Kenntnis ihrer gegenwärtigen Lage und Verfassung sowie eine gewisse Ahnung davon, was demnächst auf sie zukommen wird. Zweitens müssen Personen ihre Handlungsoptionen kennen und abschätzen können. D.h. sie müssen wissen, welche Handlungen sie gegenwärtig durchführen können und abschätzen können, welche Konsequenzen ihr Handeln nach sich ziehen wird.

Eine Situation, in der es Personen schwerfällt, ihre gegenwärtige Lage ein- und die Folgen ihrer Handlungsoptionen abschätzen zu können, bezeichnen wir als *ungewiss*. Diese *Ungewissheit* kann zustande kommen, weil es der Person an Informationen über ihre Situation fehlt oder weil sie den vorliegenden Informationen misstraut, sie als verzerrt, ungenau oder schlichtweg falsch einschätzt (Smithson 1989). Eine gegenteilige Situation hingegen, in der einer Person Informationen zur Verfügung stehen, die sie als ausreichend und verlässlich genug einstuft, um ihre aktuelle Lage einschätzen und antizipieren zu können, was auf sie zukommen wird, und um ihre Handlungsoptionen inklusive deren möglicher Folgen abschätzen zu können, beschreiben wir als *gewiss* (Dewey 2001).

Wir nutzen die Begriffe *Ungewissheit* und *Gewissheit* demnach nicht, um die objektive Qualität von Informationen zu beschreiben und Aussagen darüber zu treffen, wie richtig und verlässlich Informationen sind. Um Aussagen dieser Art zu treffen, etwa ob eine physikalische Theorie die Welt beschreibt oder eine errechnete Prognose eintrifft, greifen wir auf die Begriffe der Richtigkeit und Verlässlichkeit zurück. Dagegen nutzen wir die Begriffe *Ungewissheit* und *Gewissheit*, um das subjektive Überzeugt-sein von Informationen bzw. subjektive Überzeugungen zu beschreiben, die Personen zum Ausgangspunkt ihres Entscheidens und Handelns machen. Vor diesem Hintergrund beschreibt *Ungewissheit* einen subjektiven Mangel an Informationen, der das Treffen vernünf-

tiger Handlungsentscheidungen erschwert oder verunmöglicht (Wittgenstein 2020) – *Gewissheit* hingegen das Vorliegen von Informationen, die als verlässlich eingestuft werden und die Grundlage des eigenen Handelns und Entscheidens bilden (Dewey 2001). Diese strikte Differenzierung zwischen einer objektiven Qualität von Informationen und einem subjektiven Überzeugt-sein von Informationen bedeutet jedoch nicht, dass die Richtigkeit bzw. Verlässlichkeit einer Information für subjektive Überzeugungen bzw. Gewissheit irrelevant seien. Denn als je »richtiger« und verlässlicher sich Informationen erweisen, desto eher eignen sich Personen diese Informationen als Gewissheiten an und gründen ihre Handlungsentscheidungen auf diesen.[1]

An diesem pragmatischen Zugang ist hervorzuheben, dass es weder absolute Ungewissheit noch totale Gewissheit gibt. Denn selbst dann, wenn eine Person sich ganz unsicher ist, ihre Situation kaum einschätzen und ihre Handlungen und deren Folgen schlecht abwägen kann, besitzt sie immer noch ein Mindestmaß an Situationsbewusstsein. Sie ist, vereinfacht ausgedrückt, niemals absolut unwissend und kann immer auf einen Mindestbestand an Kenntnissen oder Intuitionen zurückgreifen, um sich und ihre Handlungen zu orientieren (Wittgenstein 2020). Umgekehrt ist es, zumindest im Regelfall, auch so, dass jede Person um die Fehlbarkeit ihrer eigenen Überzeugungen weiß. Sie weiß, dass selbst, wenn sie sich einer Sache momentan wirklich sicher ist, zukünftige Informationen ihr gegenwärtiges Wissen als defizitär, verzerrt oder fehlgeleitet erweisen können und dass zumindest die Möglichkeit besteht, dass andere Informationen zutreffender sind und eine verlässlichere Grundlage für das eigene Handeln liefern (Johnson 2022). Jede Ungewissheit und jede Gewissheit ist demnach vorläufig und relativ.

Da es uns in unserem Beitrag weniger darum geht, die gegenwärtigen Potentiale spezifischer *KI*-Technologien zu bewerten denn eher darum, zu erkunden, wohin die Reise gehen könnte und welche Gewissheiten mithilfe

1 Unsere Präferenz für diese pragmatischen Begriffsdefinition resultiert primär aus der Beobachtung, dass individuelle Entscheidungen in sozialen Kontexten häufig auf eingeschränkt verlässlichen Informationen basieren. Dennoch können solche Entscheidungen gleichzeitig als von Gewissheit getragen wahrgenommen werden. Ein exemplarisches Paradigma hierfür sind juristische Prozesse, insbesondere im Kontext der Beweisführung und richterlichen Urteilsfindung. Im Rahmen einer Beweiswürdigung können Richter*innen in ihrem Urteil eine Tatsache (z.B., dass das Fahrzeug A auf das Fahrzeug B aufgefahren ist) als mit Gewissheit gegeben aussprechen, auch wenn die dafür beweisgebende Zeug*innenaussage längst nicht alle Zweifel ausgeräumt hat (z.B., weil die Wahrnehmung der Zeug*innen schlicht zu lange zurückliegt).

von *KI* in Zukunft produziert werden könnten, greifen wir auf ein alltagssprachliches Verständnis von *KI* zurück. Unter *KI* verstehen wir sämtliche Algorithmen, die gegenwärtig unter dem Begriff *KI* subsummiert und dazu genutzt werden, große Datenmengen hochgradig selbstständig zu analysieren und Schlüsse aus ihnen zu ziehen (Boden 2018). Diskussionen darüber, ob es sich bei Algorithmen, die als *KI* betitelt werden, tatsächlich um *KI* handelt oder nur um komplexe Wenn-Dann-Schleifen, sowie feingliedrige Differenzierungen in verschiedene Arten von *KI* klammern wir an dieser Stelle bewusst aus.

Unter dem Begriff *Gesundheitssystem* fassen wir sämtliche Institutionen, die innerhalb eines meist nationalen Kontextes an der medizinischen Behandlung oder gesundheitlichen Pflege von Personen, der Organisation dieser Tätigkeiten oder ihrer Finanzierung beteiligt sind (Schölkopf und Grimmeisen 2021). Viele dieser Gesundheitssysteme weisen eine duale öffentlich-private Struktur auf. Der öffentliche Teil des Gesundheitssystems, dem zumeist alle Bürger*innen obligatorisch angehören, bietet ihnen eine medizinische Grundversorgung. Um über dieses Grundmaß hinaus medizinisch versorgt zu sein, können die Bürger*innen substitutive oder zusätzliche Versicherungen abschließen (Rice 2021). Dies bezeichnen wir als den privaten Teil des Gesundheitssystems. Im deutschen Gesundheitssystem entspricht die GKV dem öffentlichen Teil des Gesundheitssystems, während die PKV das Äquivalent zum privaten Gesundheitssystem bildet.

Kann KI dazu beitragen, gesundheitliche Gewissheiten zu schaffen?

Ehe wir der Frage nachgehen können, welche Auswirkungen es auf das öffentliche Gesundheitssystem hat, wenn KI gesundheitliche Ungewissheiten reduziert und gesundheitliche Gewissheiten produziert, müssen wir erst der Frage nachgehen, ob und wie KI überhaupt in der Lage ist bzw. sein wird, dies zu tun. Unser erster Schritt wird sein, die Begriffe *gesundheitliche Unsicherheiten* und *gesundheitliche Gewissheiten* genauer zu definieren. Danach werden wir aufzeigen, wie aktuelle KI-Anwendungen bereits gegenwärtig dazu beitragen, Unsicherheiten im Gesundheitsbereich zu mindern und bedingte Gesundheitsgewissheiten zu schaffen. Zum Schluss werden wir einen Blick in die Zukunft werfen und erwägen, wie sich das Potential von KI zur Schaffung gesundheitlicher Gewissheiten in den kommenden Jahren entwickeln könnte.

Ungewissheit haben wir als den Status bezeichnet, in dem einer Person nicht ausreichend verlässliche Informationen zur Verfügung stehen, um ihre gegenwärtige Situation einzuschätzen, zukünftige Entwicklungen zu antizipieren und auf dieser Grundlage informierte Handlungsentscheidungen zu treffen. Insbesondere in gesundheitlichen Kontexten ist diese Art von Ungewissheit äußerst präsent (Hatch 2016). Denn im Regelfall besitzen Personen nicht die nötigen Informationen über ihren Körper, um ihre gegenwärtige gesundheitliche Verfasstheit präzise einzuschätzen, geschweige denn die Entwicklung ihrer Gesundheit in den nächsten Jahren vorherzusagen. Selbst wenn sie umfangreiche Informationen haben, erschwert die hohe Komplexität des menschlichen Körpers das Ziehen klarer Schlüsse über ihre aktuelle und zukünftige Gesundheit. Diese *gesundheitliche Ungewissheit*[2] hat praktische Konsequenzen. Etwa, dass viele Menschen einen gesundheitsfördernden Lebensstil führen, um eventuellen Krankheiten vorzubeugen, oder Versicherungen abschließen, um im Falle, dass doch eine Krankheit eintritt, finanziell abgesichert zu sein. Gesundheitliche Gewissheit hingegen existiert dort, wo Personen davon überzeugt sind, ihren gegenwärtigen Gesundheitszustand und zukünftige Gesundheitsentwicklungen gut einschätzen zu können (Aronowitz 2015). Die Quellen solch gesundheitlicher Gewissheit können vielfältig sein. Häufig sind es Informationen über den eigenen Körper oder computergestützte Vorhersagen, die von den Personen als verlässlich eingestuft werden, die einer solchen Gesundheitsgewissheit zugrunde liegen.[3] Auch eine solche Gesundheitsgewissheit kann praktische Konsequenzen zeitigen. Auf einige davon werden wir im folgenden Kapitel genauer eingehen.

Wie Dewey in seinem Werk *Streben nach Gewißheit* herausarbeitet, bemühen sich Menschen seit je her, herrschende Ungewissheiten in handlungsorientierende Gewissheiten zu überführen – und nutzen dazu sämtliche wissenschaftliche und technische Möglichkeiten (Dewey 2001). Dies gilt insbesondere auch für den Gesundheitsbereich, wo die aktuellsten wissenschaftlichen sowie

2 Wenn wir von *gesundheitlicher Ungewissheit* oder *gesundheitlicher Gewissheit* sprechen, beziehen wir uns stets auf die individuelle Ungewissheit bzw. Gewissheit einer Person hinsichtlich ihrer Gesundheit.

3 An dieser Stelle ich wichtig, erneut zu betonen, dass Gewissheit nicht Richtigkeit bedeutet. Nur weil eine Person davon überzeugt ist, dass sie gesund ist und dies in Zukunft bleiben wird, d.h. ein hohes Maß an gesundheitlicher Gewissheit besitzt, muss dies nicht zutreffen. Es kann auch vorkommen, dass Personen fest vom Gegenteil dessen überzeugt sind, was »wirklich der Fall« ist – auch was ihre Gesundheit angeht.

technischen Herangehensweisen genutzt werden, um zusätzliche Informationen zu erlangen, diese auswerten zu können und dadurch einen Eindruck über die gesundheitliche Lage und Zukunft einer Person zu erlangen (Aronowitz 2015). In jüngster Zeit werden hierfür insbesondere KI-Anwendungen herangezogen (Tretter et al. 2023).

Ein eindrückliches Beispiel hierfür sind sogenannte »Digitale Zwillinge«, d.h. KI-generierte virtuelle Abbilder realer Personen. Diese basieren auf den biomedizinischen (z.B. Herzfrequenz, Sauerstoffsättigung im Blut, Blutdruck) und Lebensstildaten (z.B. Bewegung, körperliche Aktivität, Schlafzyklen, Konsummuster, Ernährung) der Repräsentierten. Diese Daten werden von Sensoren in Echtzeit aktualisiert und von einer KI modelliert, um dynamische *in-silico*-Simulationen von Personen zu erstellen. Obwohl diese Technik derzeit vor allem zu Forschungszwecken eingesetzt wird, gibt es doch eine Vielzahl von Studien, die ihren Nutzen in der medizinischen Praxis testen (Ahmadi-Assalemi u.a. 2020). So werden digitale Zwillinge verwendet, um in virtuellen Simulationen zu testen, wie gut ein*e Patient*in auf verschiedene Medikamente reagiert und welches die beste Wirksamkeit hat (Björnsson u.a. 2020); oder um virtuelle Operationen an einer Person durchzuführen und zu simulieren, ob und wie sie diese verträgt und welchen Nutzen die Operationen für sie haben (Ahmed und Devoto 2021).

Darüber hinaus versprechen digitale Zwillinge, auf Grundlage der ihnen zur Verfügung stehenden Daten und der Extrapolation vergangener und gegenwärtiger Gesundheitstrends, personalisierte und hochpräzise Einschätzungen der gesundheitlichen Entwicklung einer Person geben zu können – etwa wie hoch deren individuelles Risiko ist, binnen eines bestimmten Zeitraums bestimmte Krankheiten zu erleiden (Hafez 2020; Huang, Kim, und Schermer 2022). Sind die Prognosefähigkeiten von digitalen Zwillingen aktuell noch auf wenige Krankheiten und einen vergleichsweise kurzen Zeitraum begrenzt, ist davon auszugehen, dass schon die nächsten Generationen digitale Zwillinge ein breites Krankheitsspektrum und einen langen Zeitraum abdecken und in naher Zukunft bereits ein vollständiges Gesundheitsprofil von Einzelpersonen, inklusive all ihrer individuellen Krankheitsrisiken, erstellen werden können.

Natürlich ist nicht davon auszugehen, dass digitale Zwillinge oder vergleichbare KI-Systeme den Gesundheitsverlauf einer Person jemals mit *absoluter* Präzision prognostizieren oder ihre Krankheitsrisiken *zweifelsfrei* bestimmen können. Bereits das Vorliegen unvollständiger oder verzerrter Datensätze limitiert die Genauigkeit, die KI-Prognosen erreichen können,

maßgeblich. Doch liefern derartige KI-Systeme Erwartungswerte, die auf präzisen Daten und hochfunktionalen Modellen basieren, folglich eine große Plausibilität besitzen und den Personen, die mit ihnen interagieren – in erster Linie den Patient*innen und dem ärztlichen Personal – gesundheitliche Gewissheit verschaffen.

Wie stellen neue Gesundheitsgewissheiten die Gesundheitssysteme vor Herausforderungen?

Selbst wenn KI-Prognosen niemals unfehlbar und die entstehenden Gesundheitsgewissheiten niemals unanfechtbar sind, bergen sie doch große Chancen.[4] Wo etwa Patient*innen davon überzeugt sind, in naher Zukunft an einer Krankheit zu erkranken, wenn sie ihren gegenwärtigen Lebensstil beibehalten, kann diese Gewissheit – unabhängig davon, wie richtig oder falsch sie ist – sie zu einem gesünderen Lebenswandel motivieren und dazu beitragen, einer eventuellen Krankheit tatsächlich präventiv entgegenzuwirken. Im Idealfall können Krankheiten durch solch vorausschauende Interventionen vollständig verhindert und Personen eine Menge Leid erspart werden (Vaishya u.a. 2020). Auch für die leitenden Institutionen der Gesundheitssysteme stellen präzise KI-Prognosen eine große Chance dar. Denn wo sie sich auf Prognosen verlassen und präzise einschätzen können, wie viele Personen in Zukunft erkranken werden und medizinisch behandelt werden müssen, wird es Gesundheitssystemen besser denn zuvor möglich, weitsichtiger und sicherer zu planen (Panch, Szolovits, und Atun 2018; Schwalbe und Wahl 2020). Sie können notwendige Behandlungskapazitäten und finanzielle Ressourcen freisetzen oder schaffen und dort abziehen, wo sie nicht benötigt werden. So kann der medizinische Einsatz von KI den vorherrschenden Kostendruck abmildern und die Gesundheitssysteme unterstützen (Knorre, Müller-Peters, und Wagner 2020).[5]

---

4 Es ist zu erwarten, dass selbst Prognosen, die »nur« zu 80 Prozent verlässlich sind, oder »nur« 75 % präzise sind, Patient*innen helfen, Entscheidungen zu treffen und zu handeln. Dies gilt umso mehr, wenn man davon ausgeht, dass die Forschung zu KI in der Medizin in den kommenden Jahren und Jahrzehnten weiter so voranschreiten wird wie in den letzten Jahren (Pouly u.a. 2020) und immer noch präzisere Prädiktionen erlauben wird.

5 Zugegebenermaßen erweisen sich die Auswirkungen dieser neuen KI-Prädiktionen auf Ebene der Gesundheitssysteme nicht so einschneidend wie auf der Ebene der in-

So positiv sich das KI-gestützte Schaffen gesundheitlicher Gewissheiten auf den Gesundheitsbereich auswirken kann, so schwerwiegende Bedenken weckt es auf der anderen Seite (de Boer 2020; Tretter et al. 2021; Margetts und Dorobantu 2019; Coeckelbergh 2022). Eine sich besonders aufdrängende Frage lautet: Wie wirken sich wachsende gesundheitliche Gewissheiten auf Gesundheitssysteme und insbesondere deren finanzielle Stabilität aus?

Wir werden diese Frage am Beispiel des dualen Gesundheitssystems in Deutschland untersuchen, das sich aus der gesetzlichen Krankenversicherung (GKV) und der privaten Krankenversicherung (PKV) zusammensetzt und die Möglichkeit bietet, zwischen beiden zu wechseln. Während beide Versicherungssysteme ihren Versicherten prinzipiell eine ähnliche Absicherung gegen Krankheitsrisiken bieten, unterscheiden sie sich vor allem in der Art und Weise, wie sie ihre Prämien bzw. Beiträge berechnen. Nachdem wir diesen Unterschied erläutert und erörtert haben, welche Versicherungsform für bestimmte Personengruppen attraktiv sein könnte, werden wir darlegen, wie neu erlangte Gesundheitsgewissheiten das Phänomen der adversen Selektion verstärken können und welche Risiken dies für die Finanzierung der gesetzlichen Krankenversicherungen darstellt.

Deutschlands duales Gesundheitssystem – GKV und PKV

Die Gesundheitssysteme nahezu aller westlichen Länder sind solidarisch finanziert (Rice 2021; Schölkopf und Grimmeisen 2021), d.h. ihre Mitglieder zahlen eine regelmäßige Prämie in einen gemeinsamen Versicherungspool ein. Bei Erkrankung einer Person werden die anfallenden Kosten für Behandlung, Rehabilitation und sonstige Ausgaben aus diesem gemeinsamen Topf beglichen. Die Mitglieder unterstützen sich also durch ihre regelmäßigen Prämien solidarisch.

Das deutsche Gesundheitssystem stützt sich hauptsächlich auf die gesetzliche Krankenversicherung (GKV) und die (substitutive) private Krankenversicherung (PKV). Diese unterscheiden sich grundlegend in der Berechnung

---

dividuellen Gesundheit von Einzelpersonen. Denn gerade die Institutionen des Gesundheitswesens haben Zugriff auf umfassende statistische Gesundheits- und Krankheitsdaten. Und da auf institutioneller Ebene weniger interessiert, wer genau woran erkranken wird, sondern eher, wie viele Personen insgesamt woran erkranken werden, reichen statistische Gesundheitsdaten schon sehr weit. Nichtsdestotrotz können die durch KI präzisierten Informationsgewinne auch auf der institutionellen Ebene des Gesundheitssystems nochmal neue Planungssicherheit ermöglichen.

der individuellen Prämien. Bei der PKV richtet sich die Prämienhöhe typischerweise nach dem Krankheitsrisiko des Versicherten (Müller-Peters und Wagner 2017). Ist das Risiko hoch, steigen auch die Prämien, und umgekehrt. Dieses individuelle Krankheitsrisiko ermitteln private Krankenversicherungen anhand spezifischer statistischer und verhaltensbezogener Daten, wie beispielsweise chronische Vorerkrankungen, das Alter und ob die versicherte Person raucht (Albrecht 2018; Hoffmann 2021).

Im Gegensatz dazu folgt die Prämienberechnung in der GKV zwei Umverteilungs- oder Subventionsmechanismen. Erstens orientiert sich die Prämienhöhe nicht am individuellen Krankheitsrisiko. Das bedeutet, dass Personen mit einem hohen Risiko nicht notwendigerweise höhere Prämien zahlen als solche mit einem geringeren Krankheitsrisiko. Dieses Prinzip der »subventionierenden Risikosolidarität« (Lehtonen und Liukko 2011) soll verhindern, dass Personen mit einem hohen Krankheitsrisiko durch hohe Versicherungsprämien zusätzlich belastet werden. Zweitens richten sich die Prämien nach dem individuellen Einkommen. Personen mit einem hohen Einkommen zahlen höhere Prämien als Personen mit einem niedrigen Einkommen. Mit dieser »subventionierenden Einkommenssolidarität« (Lehtonen und Liukko 2011) wird angestrebt, dass Personen mit einem niedrigen Einkommen nicht einen überproportional hohen Teil ihres Einkommens für ihre Gesundheitsvorsorge aufwenden müssen, während Personen mit einem hohen Einkommen dafür nur einen verhältnismäßig kleinen Teil ihres Einkommens benötigen. Aktuell müssen die meisten Versicherten in der GKV mindestens 14,6 % ihres beitragspflichtigen Einkommens abführen (§ 241 SGB V). Dabei wird die Berechnungsgrundlage durch die sogenannte Beitragsbemessungsgrenze – aktuell 59.850 Euro (Bundesregierung 2023) – nach oben begrenzt (§ 223 Abs. 3 SGB V). Die Beiträge werden in Gesundheitsfonds gesammelt und anschließend nach bestimmten Risikostrukturen an die gesetzlichen Krankenkassen verteilt (§ 266 SGB V). Erhält eine Krankenkasse dabei nicht genügend Mittel, etwa weil sie Personen mit sehr hohen Krankheitsrisiken versichert, muss sie einen eigenen Zusatzbeitrag erheben (§ 242 SGB V). Es gibt jedoch keinen Regelungsrahmen, der die Höhe des Krankheitsrisikos und den einkommensabhängigen Beitrag miteinander verknüpft.

In den meisten westlichen Ländern, mit Ausnahme der Vereinigten Staaten, gibt es obligatorische öffentliche Gesundheitssysteme (Rice 2021; Schölkopf und Grimmeisen 2021). In fast allen diesen Ländern haben die Bürger*innen die Möglichkeit, *zusätzliche* private Krankenversicherungen für Leistungen abzuschließen, die nicht oder nicht vollständig durch die öffentliche Ge-

sundheitsversorgung abgedeckt werden. In einigen Ländern, darunter auch Deutschland, können Bürger*innen auch vollständig in die private Krankenversicherung wechseln. Aktuell sind in Deutschland jedoch 88,2 % der Bürger*innen in der GKV versichert und nur 5,2 % ausschließlich privat (GKV-Spitzenverband 2021).

Interessen der Versicherten

Die Teilnahme an Gesundheitssystemen bietet Versicherten den erheblichen Vorteil, im Krankheitsfall finanziell abgesichert zu sein und nicht unmittelbar für anfallende Behandlungskosten aufkommen zu müssen. Im Gegenzug entstehen den Versicherten durch regelmäßige Beitragszahlungen Kosten. Bei der Auswahl ihrer Versicherungsform nehmen viele Menschen eine Abwägung der anfallenden Kosten und der zu erwartenden Leistungen vor (Schmitz 2017; Richter, Ruß und Schelling 2019; Zerth 2020), um die Option zu wählen, die ihnen den größtmöglichen Gesamtnutzen bietet (Corea 2019). Da die Leistungen in GKV und PKV im Grunde gleich sind – beide bieten eine Grundabsicherung; die PKV bietet einige Zusatzleistungen (vgl. § 11 SGB V, § 192 VVG, § 152 VAG) (Becker und Kingreen 2022) –, sind es primär die Kosten, die die Entscheidung zwischen beiden Versicherungsformen beeinflussen (Schmitz 2017; Zerth 2020; Lünich und Starke 2021; Schmitz 2017; Zerth 2020).

Solange es nicht möglich ist, den Gesundheitsverlauf einer Einzelperson präzise zu prognostizieren und somit eine gesundheitliche Ungewissheit besteht, gleicht die Wahl zwischen GKV und PKV einer »Lotterie«. Weder die Einzelpersonen noch die Versicherungen können erahnen, wie sich die Gesundheit einer Person entwickeln wird und wie hoch die Prämien sein müssen, um diese Kosten zu decken (Jong 2021). Ein hypothetisches Szenario, in dem KI in der Lage ist, den Gesundheitsverlauf von Einzelpersonen präzise vorherzusagen und gesundheitliche Gewissheit zu schaffen, würde diese Situation drastisch verändern. Denn für die Person, der die KI einen ungünstigen Gesundheitsverlauf prognostiziert, läge es dann nahe, über die GKV versichert zu sein. Hingegen wäre es für Personen, denen eine günstige Gesundheitsentwicklung und damit ein sehr geringes Krankheitsrisiko prognostiziert wird, vorteilhaft, sich privat zu versichern und eine stärker individualisierte Prämie zu zahlen.

Adverse Selektion und Finanzierungsprobleme für öffentliche Gesundheitssysteme

Wo Einzelpersonen Gewissheit über ihre gesundheitliche Entwicklung besitzen, ermöglicht ihnen dies eine genauere Kosten-Nutzen-Bewertung bei ihrer Versicherungswahl als bisher. Diese Möglichkeit kann für Personen mit günstigen Gesundheitsprognosen und »geringem Risiko« von Vorteil sein. Für sie könnte ein Anreiz bestehen, in die PKV zu wechseln, in der die Offenlegung ihres prognostizierten Gesundheitsverlaufs bzw. ihrer geringen Krankheitsrisiken möglicherweise zu niedrigeren Prämienzahlungen führen kann. Für Personen mit ungünstigen Gesundheitsprognosen und deutlich erhöhtem Krankheitsrisiko, die bereits in der PKV sind, könnte die neue gesundheitliche Gewissheit dagegen ein Problem darstellen. Sie müssen mit einem erheblichen Anstieg ihrer individuellen Prämien rechnen, wenn auch die PKV Zugang zu ihren Gesundheitsprädiktionen erhält – selbst unter den engeren Voraussetzungen für die Prämienanpassung in der PKV nach § 203 Abs. 2 VVG.[6]

Diese (gegenwärtigen, aber vor allem zukünftigen) Möglichkeiten, den eigenen Gesundheitsverlauf präzise zu prognostizieren, können auch eine große Herausforderung für die finanzielle Stabilität der öffentlichen Gesundheitssysteme als Ganzes darstellen (van Kleef, Eijkenaar, und van Vliet 2020). Dies gilt besonders für duale Systeme wie das Deutschlands, in denen die Bürger*innen die Möglichkeit haben, aus der öffentlichen GKV auszusteigen und stattdessen eine private Krankenversicherung abzuschließen. Ein erheblicher Teil der Personen mit günstigen Gesundheitsprognosen hätte einen Anreiz, von dieser Möglichkeit Gebrauch zu machen und sich privat zu versichern, um Geld zu sparen. Umgekehrt haben Personen, denen ein negativer Gesundheitsverlauft vorausgesagt wird, in der PKV ein Interesse daran, in die GKV zu wechseln. Dieser Effekt löst eine spaltende Entwicklung aus: Personen mit schlechteren gesundheitlichen Prognosen verbleiben in der GKV oder wechseln dorthin, während Personen mit günstigen Gesundheitsprognosen in die PKV wechseln oder privat versichert bleiben, wenn sie es bereits sind.

6 Danach ist eine einseitige Prämienanpassung durch private Versicherungsunternehmen, die besonders an die Versicherungsverträge gebunden sind – was bei der substitutiven PKV regelmäßig der Fall ist –, nur möglich, wenn die wesentlichen Rechnungsgrundlagen der Sterbewahrscheinlichkeit und der Versicherungsleistungen sich nachhaltig verändern und dies durch einen unabhängigen Treuhänder bestätigt wurde.

Dieses Phänomen ist ein Fall von »adverser Selektion« (Bitter und Uphues 2017).

Im Fall der GVK ist adverse Selektion ein Problem, da sie die Allokation zwischen Personen mit positiven Gesundheitsprognosen und »geringem Risiko« und Personen mit negativen gesundheitlichen Prognosen und »hohem Risiko« allmählich in Richtung eines höheren Durchschnittsrisikos verschieben könnte, während Personen mit geringeren Risiken in die PKV abwandern. Wenn das durchschnittliche Risiko (pro versicherter Person) steigt, steigen analog die durchschnittlich erwarteten Kosten (pro versicherter Person). Da das Versicherungssystem seine Einnahmen und Ausgaben langfristig ausgeglichen halten muss (vgl. § 220 Abs. 1 SGB V), führt dies zu steigenden Beiträgen – wenn am bestehenden Leistungsumfang festgehalten werden soll. Wenn die Beiträge steigen, kann dies eine negative Abwärtsspirale in Gang setzen, da die höheren Beiträge bedeuten würden, dass nun noch mehr Personen einen Anreiz haben, aus der Versicherung auszutreten – der Kreislauf beginnt von neuem (Cutler und Reber 1998; Cutler und Zeckhauser 1998).

Im Hinblick auf die Finanzierung der GKV (Bitter und Uphues 2017) ist es außerdem wichtig zu bedenken, dass »gesündere« Personen im Durchschnitt mehr in die gesetzliche Krankenversicherung einzahlen, als sie persönlich an gedeckten Krankenbehandlungskosten zurückerhalten. Somit erhält die Krankenkasse einen Nettoeinnahmegewinn von Personen mit günstigem Gesundheitsverlauf, der es ihr ermöglicht, Personen mit negativem Gesundheitsverlauf zu subventionieren, die mehr Kosten als Einnahmen verursachen (da ihr Risiko höher ist als das, für das sie zahlen). Der Wettbewerb zwischen GKV und PKV konzentriert sich entsprechend typischerweise auf junge, alleinstehende Personen mit hohem Einkommen (Viellehner 2017). Um Einnahmeverluste aufgrund von adverser Selektion auszugleichen, muss die GKV daher die Beiträge anheben, was bereits dem Grundsatz der Beitragsstabilität (§ 71 Abs. 1 SGB V) widerspricht und außerdem den Zweck der GKV gefährdet, den allgemeinen Zugang zu medizinischer Behandlung insbesondere für arme oder schutzbedürftige Personen sicherzustellen (Prasuhn und Wilke 2021). Lässt man diese Eskalationsdynamik ungehindert fortlaufen, birgt dies die Gefahr, die Finanzierung des Gesundheitssystems substanziell zu gefährden, da die Prämien in ungeahnte Höhen steigen könnten und die Gruppe derer, die sie bezahlen können, kontinuierlich schrumpfen würde.[7] Dieser

7 Zwar existieren regulatorische Konzepte, welche die Ausrichtung von Versicherungsentscheidungen anhand von medizinischen Gewissheiten begrenzen, z.B. durch allge-

Effekt könnte schließlich den Gesetzgeber zum Eingreifen veranlassen, da die finanzielle Stabilität der GKV auch nach der Rechtsprechung des BVerfG ein erhebliches Interesse von Verfassungsrang darstellt (BVerfG, Beschl. v. 31.10.1984 – 1 BvR 35/82, 1 BvR 356/82, 1 BvR 794/82) (Schaks 2007).

Fazit

Basierend auf der Beobachtung, dass KI immer besser darin wird, detaillierte Prognosen über die gesundheitliche Entwicklung von Individuen zu erstellen, untersucht dieser Beitrag die möglichen Auswirkungen solcher Prädiktionsfortschritte auf die finanzielle Stabilität von Gesundheitssystemen. Denn in einem zukünftigen Szenario, in dem KI hochpräzise und äußerst verlässliche Vorhersagen über den individuellen Gesundheitsverlauf treffen kann, erlangen Einzelne eine Art subjektive Gewissheit über ihren Gesundheitszustand. Diese Gewissheit können sie verwenden, um ihre Wahl zwischen GKV und PKV so optimal und kosteneffizient wie möglich zu gestalten. Für Teilnehmer*innen an öffentlichen Gesundheitssystemen spielen beitragsentscheidende Faktoren schon jetzt eine erhebliche Rolle (Schmitz 2017; Zerth 2020).

Der Unterschied in der Konzeption der PKV, die sich auf risikobasierte Versicherungsprämien stützt, und der GKV, welche einkommensbasierte Prämien verwendet, könnte klare Anreize schaffen, dass Personen mit günstigen Gesundheitsprognosen und niedrigen Krankheitsrisiken von der GKV zur PKV wechseln. Entscheidend für diesen Wechsel ist der Vergleich der Kosten, die diese Personen für die Absicherung ihres geringen Krankheitsrisikos in der GKV und PKV aufbringen müssten. Bei den einkommensabhängigen GKV-Prämien würde die Person einen Betrag x zahlen (14,6 % des beitragspflichtigen Einkommens bis maximal zur Beitragsbemessungsgrenze von 59.850 Euro pro Jahr für 2023, zzgl. kassenindividueller Zusatzbeiträge), bei den risikobasierten PKV-Prämien einen Betrag y (bemessen anhand definierter Risikobemessungsgrundlagen). Solange y geringer als x ist, besteht zumindest ein klarer finanzieller Anreiz zum Wechsel. Tendenziell könnten daher Personen mit hohem Einkommen und günstigen Gesundheitsprognosen dazu nei-

---

meine Begrenzung der Wahlmöglichkeit zwischen GKV und PKV über Versicherungsfreiheit oder durch die Beschränkung der Prämienberechnung in der PKV auf spezifische Rechengrößen (Brömmelmeyer 2017). Diese können den Effekt einer adversen Selektion jedoch nicht nachhaltig eingrenzen.

gen, die GKV zu verlassen und sich in der PKV zu versichern. Im Gegensatz dazu hätten Personen mit negativen Gesundheitsprognosen, bei denen x geringer als y ist, einen Anreiz, die einkommensabhängigen Prämien der GKV zu zahlen. Insgesamt könnte dies dazu führen, dass das Beitragsaufkommen der GKV erheblich sinkt, während die Ausgaben steigen, da sich hohe Krankheitsrisiken in der GKV ansammeln. Dieses Phänomen der adversen Selektion könnte die finanzielle Stabilität der GKV nachhaltig bedrohen.

Finanzierung

Dieser Beitrag wurde durch Zuschüsse des Bundesministeriums für Forschung und Bildung finanziert (Förderkennzeichen: 01GP1905A, 01GP1905B, 01GP1905C, 01GP2202B).

Literatur

Ahmadi-Assalemi, Gabriela, Haider Al-Khateeb, Carsten Maple, Gregory Epiphaniou, Zhraa A. Alhaboby, Sultan Alkaabi, und Doaa Alhaboby. 2020. »Digital Twins for Precision Healthcare«. In *Cyber Defence in the Age of AI, Smart Societies and Augmented Humanity*, herausgegeben von Hamid Jahankhani, Stefan Kendzierskyj, Nishan Chelvachandran, und Jaime Ibarra, 133–158. Advanced Sciences and Technologies for Security Applications. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-35746-7_8.

Ahmed, Hanad, und Laurence Devoto. 2021. »The Potential of a Digital Twin in Surgery«. *Surgical Innovation* 28 (4): 509–510. https://doi.org/10.1177/1553350620975896.

Albrecht, Peter. 2018. »Tarifierung in der Privatversicherung: Big Data, Risikoadäquanz, Solidarität«. *Zeitschrift für die gesamte Versicherungswissenschaft* 107 (5): 449–467. https://doi.org/10.1007/s12297-018-0409-2.

Aronowitz, Robert. 2015. *Risky Medicine: Our Quest to Cure Fear and Uncertainty*. Chicago: University of Chicago Press.

Becker, Ulrich, und Thorsten Kingreen. 2022. »§ 11 SGB V – Leistungsarten, Rn. 19–21«. In *SGB V – Gesetzliche Krankenversicherung*, herausgegeben von Ulrich Becker und Thorsten Kingreen, 8. Aufl. München: C.H. Beck.

Bitter, Philip, und Steffen Uphues. 2017. »Big Data und die Versichertengemeinschaft – ›Entsolidarisierung‹ durch Digitalisierung?« Zuletzt zugegriffen am 31. Juli 2023. https://www.abida.de/sites/default/files/13%20Entsolidarisierung.pdf.

Björnsson, Bergthor, Carl Borrebaeck, Nils Elander, Thomas Gasslander, Danuta R. Gawel, Mika Gustafsson, Rebecka Jörnsten, et al. 2020. »Digital Twins to Personalize Medicine«. *Genome Medicine* 12 (1): 4. https://doi.org/10.1186/s13073-019-0701-3.

Boden, Margaret A. 2018. *Artificial Intelligence: A Very Short Introduction*. Oxford: Oxford University Press.

Boer, Bas de. 2020. »Experiencing Objectified Health: Turning the Body into an Object of Attention«. *Medicine, Health Care and Philosophy* 23 (3): 401–411. https://doi.org/10.1007/s11019-020-09949-0.

Brömmelmeyer, Christoph. 2017. »Belohnungen für gesundheitsbewusstes Verhalten in der Lebens- und Berufsunfähigkeitsversicherung? Rechtliche Rahmenbedingungen für Vitalitäts-Tarife.« *recht und schaden* 5: 225–232.

Bundesregierung. 2023. »Änderungen der Beitragsbemessungsgrenzen 2023«. *Die Bundesregierung informiert*. Zuletzt zugegriffen am 31. Juli 2023. https://www.bundesregierung.de/breg-de/aktuelles/beitragsbemessungsgrenzen-2023-2133570.

Chaari, Lotfi. 2019. *Digital Health Approach for Predictive, Preventive, Personalised and Participatory Medicine*. Cham: Springer Nature.

Coeckelbergh, Mark. 2022. *The political philosophy of AI: an introduction*. Cambridge: Polity.

Corea, Francesco. 2019. *Applied Artificial Intelligence: Where AI can be used in Business*. Cham: Springer.

Cutler, David M., und S. J. Reber. 1998. »Paying for Health Insurance: The Trade-Off between Competition and Adverse Selection«. *The Quarterly Journal of Economics* 113 (2): 433–466. https://doi.org/10.1162/003355398555649.

Cutler, David M., und Richard J. Zeckhauser. 1998. »Adverse Selection in Health Insurance«. *Forum for Health Economics & Policy* 1 (1). https://doi.org/10.2202/1558-9544.1056.

Dabrock, Peter. »So Lässt Sich ChatGPT Verantworten.« *Spiegel*, 30. Januar 2023. https://www.spiegel.de/netzwelt/chatgpt-so-laesst-sich-kuenstliche-intelligenz-verantworten-gastbeitrag-a-d89746ff-a263-4a70-a6d2-7029bb45b7ac.

Dewey, John. 2001. *Die Suche nach Gewißheit. Eine Untersuchung des Verhältnisses von Erkenntnis und Handeln*. Übersetzt von Martin Suhr. Frankfurt a.M.: Suhrkamp.

GKV-Spitzenverband. 2021. »Zahlen und Grafiken – GKV-Spitzenverband«. *GKV-Spitzenverband*. 2021. Zuletzt zugegriffen am 31. Juli 2023. https://www.gkv-spitzenverband.de/service/zahlen_und_grafiken/zahlen_und_grafiken.jsp.

Hafez, Wael. 2020. »Human Digital Twin: Enabling Human-Multi Smart Machines Collaboration«. In *Intelligent Systems and Applications*, herausgegeben von Yaxin Bi, Rahul Bhatia, und Supriya Kapoor, 981–993. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-29513-4_72.

Hatch, Steven. 2016. *Snowball in a Blizzard. A Physician's Notes on Uncertainty in Medicine*. New York: Basic Books.

Hoffmann, Patricia. 2021. *Telematik-Tarife in der privaten Krankenversicherung: Möglichkeiten der vitaldatenbasierten Tarif-, Prämien- und Vertragsgestaltung*. Karlsruhe: Verlag Versicherungswirtschaft.

Huang, Pei-hua, Ki-hun Kim, und Maartje Schermer. 2022. »Ethical Issues of Digital Twins for Personalized Health Care Service: Preliminary Mapping Study«. *Journal of Medical Internet Research* 24 (1): e33081. https://doi.org/10.2196/33081.

Johnson, L. Syd M. 2022. *The Ethics of Uncertainty. Entangled Ethical and Epistemic Risks in Disorders of Consciousness*. New York: Oxford University Press.

Jong, Casper H. 2021. »Risk Classification and the Balance of Information in Insurance; an Alternative Interpretation of the Evidence«. *Risk Management and Insurance Review* 24 (4): 445–461. https://doi.org/10.1111/rmir.12198.

Kleef, Richard C. van, Frank Eijkenaar, und René C. J. A. van Vliet. 2020. »Selection Incentives for Health Insurers in the Presence of Sophisticated Risk Adjustment«. *Medical Care Research and Review* 77 (6): 584–595. https://doi.org/10.1177/1077558719825982.

Knorre, Susanne, Horst Müller-Peters, und Fred Wagner. 2020. *Die Big-Data-Debatte: Chancen und Risiken der digital vernetzten Gesellschaft*. Wiesbaden: Springer Fachmedien.

Lehtonen, Turo-Kimmo, und Jyri Liukko. 2011. »The Forms and Limits of Insurance Solidarity«. *Journal of Business Ethics* 103 (1): 33–44. https://doi.org/10.1007/s10551-012-1221-x.

Lünich, Marco, und Christopher Starke. 2021. »Big Data = Big Trouble for Universal Healthcare? The Effects of Individualized Health Insurance on Solidarity«. *Preprint. SocArXiv*. https://doi.org/10.31235/osf.io/3f2xs.

Margetts, Helen, und Cosmina Dorobantu. 2019. »Rethink Government with AI«. *Nature* 568 (7751): 163–165. https://doi.org/10.1038/d41586-019-01099-5.

Müller-Peters, Horst, und Fred Wagner. 2017. *Geschäft oder Gewissen? vom Auszug der Versicherung aus der Solidargemeinschaft*. Goslar: Goslar Institut.

Panch, Trishan, Peter Szolovits, und Rifat Atun. 2018. »Artificial Intelligence, Machine Learning and Health Systems«. *Journal of Global Health* 8 (2): 020303. https://doi.org/10.7189/jogh.08.020303.

Pouly, Marc, Thomas Koller, Philippe Gottfrois, und Simone Lionetti. 2020. »Künstliche Intelligenz in der Bildanalyse – Grundlagen und neue Entwicklungen«. *Der Hautarzt* 71 (9): 660–668. https://doi.org/10.1007/s00105-020-04663-7.

Prasuhn, Armin Marek, und Christina Benita Wilke. 2021. *Reformoption Bürgerversicherung? eine Nutzwertanalyse vor dem Hintergrund aktueller und künftiger Herausforderungen des deutschen Krankenversicherungssystems*. Zuletzt zugegriffen am 31. Juli 2023. https://www.fom.de/fileadmin/fom/forschung/KCV/FOM-KCV-Schriftenreihe-Band-2-Prasuhn-Wilke-Reformoption-Buergerversicherung-2021.pdf.

Rice, Thomas. 2021. *Health insurance systems: an international comparison*. London: Academic Press.

Richter, Andreas, Jochen Ruß, und Stefan Schelling. 2019. »Insurance Customer Behavior: Lessons from Behavioral Economics«. *Risk Management and Insurance Review* 22 (2): 183–205. https://doi.org/10.1111/rmir.12121.

Schaks, Nils. 2007. *Der Grundsatz der finanziellen Stabilität der gesetzlichen Krankenversicherung: eine verfassungs- und sozialrechtliche Untersuchung. Schriften zum Gesundheitsrecht, Bd. 7*. Berlin: Duncker & Humblot.

Schmidt, Jan C. 2022. »Wandel Und Kontinuität Von Wissenschaft Durch Ki. Zur Aktuellen Veränderung Des Wissenschafts- Und Technikverständnisses.« In *Künstliche Intelligenz in der Forschung: Neue Möglichkeiten und Herausforderungen für die Wissenschaft*, herausgegeben von Carl Friedrich Gethmann, Peter Buxmann, Julia Distelrath, et al., 79–125. Berlin, Heidelberg: Springer.

Schmitz, Hendrik. 2017. »Preis, Service oder Leistungen: Was beeinflusst besonders die Krankenkassenwahl von gesetzlich Versicherten?« In *Krankenversicherung im Rating*, herausgegeben von Thomas Adolph, Oliver Everling, und Marco Metzler, 279–295. Wiesbaden: Gabler Verlag.

Schölkopf, Martin, und Simone Grimmeisen. 2021. *Das Gesundheitswesen im internationalen Vergleich Gesundheitssystemvergleich, Länderberichte und europäi-*

sche Gesundheitspolitik. 4. Aufl. Berlin: Medizinisch Wissenschaftliche Verlagsgesellschaft.

Schwalbe, Nina, und Brian Wahl. 2020. »Artificial Intelligence and the Future of Global Health«. *The Lancet* 395 (10236): 1579–1586. https://doi.org/10.1016/S0140-6736(20)30226-9.

Smithson, Michael. 1989. *Ignorance and Uncertainty: Emerging Paradigms*. New York: Springer New York.

Topol, Eric J. 2020. *Deep Medicine. Künstliche Intelligenz in der Medizin. Wie KI das Gesundheitswesen menschlicher macht*. Übersetzt von Guido Lenz. Frechen: mitp.

Tretter, Max. 2021. »Perspectives on Digital Twins and the (Im)Possibilities of Control«. *Journal of Medical Ethics* 47 (6): 410–411. https://doi.org/10.1136/medethics-2021-107460.

Tretter, Max. 2023a. »Ambivalenzen gegenwärtiger Gewissheitsbestrebungen – Menschliche Entscheidungsfreiheit in einer gewisserwerdenden Welt«. In *Alexa, wie hast du's mit der Religion? Interreligiöse Zugänge zu Technik und künstlicher Intelligenz*, herausgegeben von Anna Puzio, Nicole Kunkel, und Hendrik Klinge. 135–157. Darmstadt: wbg.

Tretter, Max, Tabea Ott, und Peter Dabrock. 2023. »AI-produced certainties in health care: current and future challenges«. AI and Ethics. https://doi.org/10.1007/s43681-023-00374-6.

Valle-Cruz, David, und Rodrigo Sandoval-Almazan. »Towards an Understanding of Artificial Intelligence in Government.« *19th Annual International Conference on Digital Government Research: Governance in the Data Age, Delft, Association for Computing Machinery, 2018*.

Valle-Cruz, David, Edgar Alejandro Ruvalcaba-Gomez, Rodrigo Sandoval-Almazan, und J. Ignacio Criado. »A Review of Artificial Intelligence in Government and Its Potential from a Public Policy Perspective.« *20th Annual International Conference on Digital Government Research, Dubai, Association for Computing Machinery, 2019*.

Vaishya, Raju, Mohd Javaid, Ibrahim Haleem Khan, und Abid Haleem. 2020. »Artificial Intelligence (AI) Applications for COVID-19 Pandemic«. *Diabetes & Metabolic Syndrome: Clinical Research & Reviews* 14 (4): 337–339. https://doi.org/10.1016/j.dsx.2020.04.012.

Viellehner, Simone. 2017. *Das Finanzierungsrecht der Gesetzlichen und Privaten Krankenversicherung: Herausforderung für die Zusammenführung in ein einheitliches System*. Baden-Baden: Nomos.

Wittgenstein, Ludwig. 2020. *Über Gewißheit*. Herausgegeben von G. E. M. Anscombe und Georg Henrik von Wright. 15. Aufl. Frankfurt a.M.: Suhrkamp.

Zerth, Jürgen. 2020. »Krankenversicherungen als Agenten und Akteure in einem Gesundheitssystem der Zukunft: Eine Rollenbetrachtung zwischen Wettbewerb und Regulierung«. In *Handbuch Gesundheitssoziologie*, herausgegeben von Peter Kriwy und Monika Jungbauer-Gans, 701–722. Wiesbaden: Springer Fachmedien.

Algorithmische Differenzierung und Diskriminierung aus Sicht der Menschenwürde

Carsten Orwat

1. Einleitung

Die Effizienz von Entscheidungen zur Differenzierung von Personen soll mit Anwendungen der Künstlichen Intelligenz (KI), algorithmischen Differenzierungen und automatisierten Entscheidungssystemen (AES) verbessert werden. Allerdings sind damit auch neue Risiken für die Grundrechte verbunden, einschließlich des Risikos von Diskriminierungen und Verletzungen der Menschenwürde. Das Antidiskriminierungsrecht soll nicht nur Gerechtigkeit und Gleichbehandlung sicherstellen, sondern auch die Möglichkeiten der freien Entfaltung der Persönlichkeit schaffen und die Menschenwürde schützen. Der Beitrag beleuchtet daher Entwicklungen der KI und von algorithmischen Differenzierungen aus der Perspektive der Menschenwürde.

Algorithmen werden zunehmend zur Unterstützung und automatischen Durchführung von Entscheidungen über die Differenzierung von Menschen eingesetzt, bei so unterschiedlichen Anwendungsfeldern wie z.B. Kreditvergabe, Wohnungssuche, Sozialleistungsbemessung oder Gerichtsentscheidungen. Solche Differenzierungen wirken sich dann auf die Verfügbarkeit und Verteilung von Produkten, Diensten, Positionen, Chancen, Vorteilen oder Belastungen aus, die essentiell für die Persönlichkeitsentfaltung und Realisierungen von Autonomie und Freiheit sind. Hier kommen dann Anwendungen zum Einsatz, bei denen maschinelles Lernen als Analysemethode bei Data Mining, Profiling oder Predictive Analytics angewandt wird und die Ergebnisse der Analysemethoden als Modelle bzw. Algorithmen in Entscheidungsprozessen eingesetzt werden. Dies geschieht entweder als Unterstützung und Beratung menschlicher Entscheidender oder als (voll-)automatisierte Entscheidungssysteme.

2. Algorithmische Differenzierung und Diskriminierung

2.1 Ursachen und Formen von algorithmischen Bias

Die Ursachen von Bias bzw. Verzerrungen bei der Verwendung von Algorithmen, insbesondere des maschinellen Lernens, sind vielfältig. Sie ergeben sich aus menschlichen Entscheidungen über die verwendete Datenbasis, die Entwicklung, den Einsatz oder die Anpassung von Algorithmen. Zu den am häufigsten genannten Ursachen von Bias gehören die mit historischen Ungleichheiten und Ungleichbehandlungen kontaminierte Trainingsdatensätze, die Auswahl ungeeigneter Messgrößen bzw. Label, ungeeignet entschiedene technische Trade-offs oder die Anwendung von Algorithmen in Bereichen, für die sie nicht trainiert bzw. optimiert wurden (z.B. Mehrabi et al. 2021, Pessach und Shmueli 2022).

Bias können dazu führen, dass Produkte und Dienste für verschiedene Bevölkerungsgruppen unterschiedlich gut funktionieren. Oder algorithmische Modelle, die bei Entscheidungen über die Differenzierung von Personen über Zugänge und Verteilung von Gütern, Positionen oder Freiheiten eingesetzt werden, führen zu Ungleichbehandlungen bei Betroffenen. Ein weiteres Problem sind Stereotypisierungen bei Generativer KI (wie z.B. Chat-GPT), vor allem wenn Entscheidungen auf Grundlage ihrer Ergebnisse getroffen werden (z.B. automatisiert erstellte Zusammenfassung von Bewerbungen oder Anträgen). Mittlerweile finden sich zahlreiche Beispiele für Bias und Diskriminierungsrisiken durch Algorithmen, von denen hier nur einige wenige angeführt werden können, etwa bei Gesichtsanalysen (Buolamwini und Gebru 2018), bei der Zuweisung zu bestimmten Therapien in der Medizin (Obermeyer et al. 2019), bei KI-basierten Analysen von Videointerviews im Personalbereich (Köchling et al. 2021) oder für Stereotypisierung bei der Sprachverarbeitung (Bender et al. 2021) (weitere Beispiele in AIAAIC Repository oder Orwat 2019, Kapitel 4).

2.2 Technische Lösungen und Restrisiken

Als Reaktion auf algorithmische Diskriminierungsrisiken werden in Wissenschaft, Forschung und Praxis große Mühen darauf gerichtet, Algorithmen diskriminierungsärmer bzw. »fairer« zu gestalten. Dabei wird vor allem versucht, Bias in Datensätzen zu beseitigen und Modelle zu optimieren. Ferner sind zahlreiche mathematische Fairness-Definitionen bzw. Fairness-

Metriken entwickelt worden, um Ungleichbehandlungen quantitativ auszudrücken und um damit Systeme zu optimieren und zu vergleichen (z.B. Pessach und Shmueli 2022). Allerdings sind diese in ihren Konsequenzen für gesellschaftliche Gerechtigkeit und Anti-Diskriminierung bisher noch wenig in der Öffentlichkeit diskutiert worden.

Des Weiteren sind die technischen Lösungen des »debiasing« in mehrfacher Weise begrenzt. So sind verschiedene Fairness-Metriken nicht gleichzeitig erfüllbar. Entwickelnde und Anbietende müssen verschiedene Trade-offs entscheiden, insbesondere ob und welche Fairness-Definitionen verwendet werden, welche Grenzwerte als Restrisiken für akzeptabel gehalten werden, ebenso sind Abwägungen zwischen einzelnen Zielen und Metriken zu treffen, wie z.B. zwischen Genauigkeit der Ermittlung der Differenzierungsziele und Diskriminierungsrisiken (z.B. Mehrabi et al. 2021, Pessach und Shmueli 2022).

Einige Fairness-Metriken sind Verhältniskennziffern, die aus den Fehlerraten »Falsch-Negative« und »Falsch-Positive« und Raten von richtig erkannten Klassifizierungen mit Bezug zu bestimmten Bevölkerungsmerkmalen gebildet werden. Bisher ist jedoch noch unklar, wie gesellschaftlich mit diesen Fairness-Metriken umgegangen wird, vor allem hinsichtlich der Niveaus an zu akzeptierenden Diskriminierungsrisiken für die Gesellschaft, etwa wer sie festlegt oder in welcher Höhe sie festgelegt werden. Nach Meinung der Europäischen Agentur für Menschenrechte kann, auch beim Vorliegen von Fairness-Metriken, nur von Fall zu Fall entschieden werden, wann eine ausreichend signifikante Diskriminierung vorliegt, nicht aber anhand der Bestimmung eines abstrakten Grenzwertes (FRA 2022, 25).

Im Entwurf der KI-Verordnung der Europäischen Union (2021/0106 (COM); KI-VO-E) wird zwar auf Parameter und Metriken verwiesen, aber nicht geklärt, wer akzeptable Fehlerraten und Grenzwerte festlegt. Zudem ermöglichen Formulierungen wie »Restrisiko« oder »akzeptabel« bzw. »vertretbar« (Artikel 9 (4) KI-VO-E) oder »angemessenes Maß an Genauigkeit« (Artikel 15 (1) KI-VO-E), dass Risikovermeidung mit Wirtschaftlichkeitsüberlegungen abgewogen werden kann. Dies ist vor dem Hintergrund zu sehen, dass für das Testen der KI-Systeme, die Risikovermeidung und andere regulatorische Pflichten Kosten entstehen. Der Entwurf der KI-VO lässt offen, ob die Europäische Kommission, Standardisierungsorganisationen, Entwickelnde, Anbietende, Anwendende oder Einrichtungen, die die »compliance« zertifizieren, die normativen Entscheidungen über Diskriminierungsniveaus und damit über die gesellschaftliche Realisierung von Gerechtigkeit tref-

fen. Dies kann dazu führen, dass Restrisiken von Diskriminierungen auf gesellschaftlich nicht festgelegtem Niveau wahrscheinlich sind.

Im Gegensatz zu menschlichen Entscheidenden, bei denen eher mit punktuellen Diskriminierungen zu rechnen ist, können einzelne AES und KI-Systeme große Reichweiten haben (z.B. durch Marktkonzentration oder bei Verwendung eines Systems in vielen Verwaltungen). Trotz scheinbar geringen Fehlerraten kann dies zu systematischen Diskriminierungen führen. Damit stellt sich die Frage, ob die Verminderung von Diskriminierungen oder Vorurteilen, die gegenüber menschlichen Entscheidenden mit Algorithmen angestrebt wird, nicht wieder durch den Reichweiteneffekt der Restrisiken aufgewogen oder sogar überwogen werden können.

2.3 Restrisiken und Antidiskriminierungsrecht

Die Restrisiken müssen im Lichte der Diskriminierungsverbote des Antidiskriminierungsrechts betrachtet werden. Zu rechtlich definierten Diskriminierungen führen Bias-Probleme, wenn bei Differenzierungen verzerrende Algorithmen verwendet werden und es dadurch zu ungerechtfertigten Ungleichbehandlungen durch die Nutzung von rechtlich geschützten Merkmalen (z.B. Geschlecht, ethnische Herkunft, Religion, Behinderung, Alter oder sexuelle Identität) oder scheinbar neutralen Merkmalen, Verfahren, Vorschriften oder Praktiken, die aber einen Zusammenhang zu den geschützten Merkmalen haben, kommt (Hacker 2018, von Ungern-Sternberg 2022).

Das Antidiskriminierungsrecht zeigt jedoch Schwächen im Umgang mit algorithmischen Diskriminierungen, denn angesichts algorithmischer, oft personalisierter oder individualisierter, Differenzierungen kann es für einzelne Betroffene schwierig sein, eine Ungleichbehandlung im Verhältnis zu anderen Betroffenen als solche wahrzunehmen und die rechtlich notwendigen ersten Nachweise für eine Schlechterstellung gegenüber vergleichbaren anderen Personen zu erbringen. Diese sind jedoch Voraussetzung dafür, dass rechtliche Verfahren eingeleitet werden können, auch wenn dann die einer Diskriminierung beschuldigte Person oder Einrichtung die Beweislast hat, dass sie nicht diskriminiert (Orwat 2019, 107–09 m.w.N., von Ungern-Sternberg 2022, 1141f.). Die ohnehin hohen Hürden für betroffene Individuen, die oft verhindern, dass es zu einem rechtlichen Diskriminierungsfall kommt, werden so noch weiter erhöht.

3. Verständnis der Menschenwürde im Verfassungsrecht

Die Menschenwürde gilt oft als ein abstraktes und unterschiedlich interpretierbares Konzept (z.B. Mahlmann 2008). Dies wird oft kritisch gegen ihre Verwendung eingewandt. Allerdings ist die Menschenwürde in vielen Menschen- und Grundrechtsdokumenten enthalten und durch Umsetzung in Rechtssystemen und der Rechtsprechung konkretisiert worden (z.B. McCrudden 2008, Mahlmann 2012). Im Folgenden wird daher Bezug auf die Entscheidungen des Bundesverfassungsgerichts (BVerfGE) genommen.

Danach ist die Menschenwürde ein »fundamentaler Wert- und Achtungsanspruch […], der jedem Menschen zukommt.« (BVerfGE 87, 209, Rn. 109). Sie umfasst vor allem die Wahrung der personalen Individualität, Identität und Integrität sowie die elementare Rechtsgleichheit (BVerfGE 144, 20, Leitsatz 3 a, Rn. 539). »Dem liegt eine Vorstellung vom Menschen zugrunde, die diesen als Person begreift, die in Freiheit über sich selbst bestimmen und ihr Schicksal eigenverantwortlich gestalten kann« (ebd., Rn. 539). Die Menschenwürde ist dem Menschen inhärent. »Jeder besitzt sie, ohne Rücksicht auf seine Eigenschaften, seine Leistungen und seinen sozialen Status.« (BVerfGE 87, 209, Rn. 107).

Zur weiteren Konkretisierung der Menschenwürde wurde die so genannte »Objektformel« entwickelt (Hong 2019, 672–90). Danach ist es mit der Menschenwürde unvereinbar, den Menschen »zum bloßen Objekt« staatlichen Handelns zu machen (BVerfGE 27, 1, S. 6 und weitere Entscheidungen). Nach der Objektformel darf der Mensch nicht wie eine Sache behandelt, verdinglicht oder zu einem bloßen Gegenstand herabgewürdigt werden (Hong 2019, 418f. m.w.N.). Das Bundesverfassungsgericht hat die Objektformel mit der Subjektformel weiterentwickelt, wonach es untersagt ist, einzelne Menschen einer Behandlung auszusetzen, die ihre Subjektqualität grundsätzlich in Frage stellt, indem sie die Achtung des Wertes vermissen lässt, der ihnen um ihrer selbst willen zukommt (nach Hong 2022, Rn. 26 m.w.N., ausführlich in Hong 2019, 421–28). Nach Höfling (2021, Rn. 16) ist zur Bestimmung einer Verletzung der Menschenwürde in konkreten Entscheidungssituationen zu fragen, ob der Subjektstatus eines Menschen trotz Verobjektivierung in Unterordnungs- und Abhängigkeitsverhältnissen durch Kompensationsmechanismen noch gesichert ist.

Bestimmte Formen von Diskriminierung stellen direkt eine Verletzung der Menschenwürde dar. Ein Verstoß gegen Art. 1 (1) Grundgesetz GG wird unter anderem bei einer unmittelbaren Diskriminierung durch Eingriffe in die je-

dem zustehenden Freiheitsgrundrechte wegen der in Art. 3 (3) GG genannten Kriterien gesehen (Herdegen 2022, Rn. 120). Höfling u.a. sehen eine verbotene Menschenwürdeverletzung vor allem in rassistischer Diskriminierung und ähnlichen demütigen Ungleichbehandlungen (Höfling 2021, Rn. 35) (siehe dazu vor allem auch BVerfGE 144, 20, Rn. 541). Hillgruber betont, dass eine Verletzung der Menschenwürde nicht nur vorliegt, wenn Menschen bestimmter »Rasse«, Hautfarbe, Religion oder Geschlecht als »minderwertig« angesehen werden, sondern auch bei Diskriminierungen von Menschen aufgrund einer körperlichen und geistigen Behinderung, insbesondere wenn eine Ausgrenzung wegen ihrer Behinderung droht (Hillgruber 2023, Rn. 17). Die genannten Merkmale sind besonders relevant, da sie zum einen unveränderliche und persönlichkeitskonstituierende Merkmale sind. Zum anderen sind sie historisch als Abgrenzung von den Gräueltaten des nationalsozialistischen Unrechtsregimes begründet (Lehner 2013, 226–48, Hong 2019, 407).

Eine weitere Konkretisierung der Menschenwürde erfolgt durch ihre Weiterentwicklung zum verfassungsrechtlichen allgemeinen Persönlichkeitsrecht, aus dem vor allem das zur weiteren Realisierung des Menschenwürdeschutzes entwickelte Recht auf informationelle Selbstbestimmung, die Diskriminierungsverbote, aber auch das Recht auf Selbstdarstellung (Britz 2007) für die nachfolgenden Betrachtungen relevant sind. Das Antidiskriminierungsrecht dient so nicht nur der Realisierung des Rechts auf Gleichbehandlung und sozialpolitischer Ziele, sondern auch des Rechts auf freie Entfaltung der Persönlichkeit und dem Schutz der Menschenwürde (ähnlich Baer 2009). Das Recht auf informationelle Selbstbestimmung und die Diskriminierungsverbote dienen u.a. dazu, unangemessene Fremdbilder der Persönlichkeit zu verhindern. Die Rechte sollen dazu befähigen, mitzuentscheiden, was Betroffene als zu ihrer Persönlichkeit gehörig und diese ausmachend ansehen können (Britz 2007). Kerngewährleistung des Persönlichkeitsrechts ist es, »Mechanismen zur Verfügung zu stellen, die den Einzelnen so in die Vorgänge der Konstituierung von Persönlichkeit einbinden, dass er seine Persönlichkeit als frei gewählt begreifen kann […]« (Britz 2008, 191).

In einer Reihe von Entscheidungen hat das Bundesverfassungsgericht das Recht auf Menschenwürde und das Persönlichkeitsrecht konkret auf die Risiken der modernen Informations- und Kommunikationstechnologien bezogen und entwickelt. In der »Mikrozensus«-Entscheidung (BVerfGE 27, 1) hat das Gericht deutlich gemacht, dass es der Menschenwürde widerspricht, den Menschen zu einem bloßen Objekt im Staat zu machen. Es ist mit der Menschenwürde unvereinbar, den Menschen in seiner gesamten Persönlich-

keit zwangsweise zu registrieren und zu katalogisieren und ihn damit wie eine Sache zu behandeln, die einer Bestandsaufnahme in jeder Beziehung zugänglich ist (ebd., S. 6–7).

In der Entscheidung zur Volkszählung (BVerfGE 65, 1) wurde das Grundrecht auf informationelle Selbstbestimmung begründet, das dem Recht auf freie Entfaltung der Persönlichkeit in Verbindung mit dem Recht auf Menschenwürde dient. Es garantiert die Befugnis, grundsätzlich selbst über die Preisgabe und Verwendung seiner persönlichen Daten bestimmen zu können (ebd., Leitsatz 1). Das Recht auf informationelle Selbstbestimmung dient nicht nur der Sicherung der äußeren und inneren Freiheitsdimensionen (Handlungsfreiheit und Identitätsbildung), sondern auch der Vermeidung von Abschreckungseffekten (chilling effects), die durch Unsicherheiten über die Datenverarbeitungen bei den Betroffenen entstehen können (Britz 2010).

In der Entscheidung zum großen Lauschangriff (BVerfGE 109, 279) erkennt das Gericht im Hinblick auf die Unantastbarkeit der Menschenwürde einen Kernbereich privater Lebensgestaltung an, der absoluten Schutz genießt. Was zum Kernbereich privater Lebensgestaltung gehört, hängt davon ab, ob der Sachverhalt einen inhaltlich höchstpersönlichen Charakter hat (ebd., Rn. 123). Ebenso verletzt eine zeitlich und räumliche »Rundumüberwachung« die Menschenwürde, wenn sie über einen längeren Zeitraum erfolgt und alle Bewegungen und Lebensäußerungen des Betroffenen nahezu lückenlos erfasst und zur Grundlage eines Persönlichkeitsprofils werden können (ebd., Rn. 150).

In der Entscheidung zum Recht auf Vergessen I (BVerfGE 152, 152) legt das Bundesverfassungsgericht das Recht auf informationelle Selbstbestimmung im privaten Bereich so aus, dass dem Einzelnen gewährleistet ist, »über der eigenen Person geltende Zuschreibungen selbst substantiell mitzuentscheiden.« (Ebd.). Das Gericht sah, dass in vielen Lebenssituationen private Unternehmen die grundlegenden Dienstleistungen erbringen, die eine entscheidende Rolle bei der öffentlichen Meinungsbildung, der Zuteilung oder Verweigerung von Chancen oder der Ermöglichung der Teilhabe am sozialen bzw. täglichen Leben spielen. In vielen Fällen geschehe dies auf der Grundlage einer umfassenden Datenerhebung und Verarbeitung durch oft marktmächtige Unternehmen, bei der die Offenlegung personenbezogener Daten in großem Umfang kaum zu vermeiden ist, will man nicht von den Produkten und Diensten ausgeschlossen werden (ebd., Rn. 85). In Fällen bei weitreichenden Abhängigkeiten oder Ausgesetztsein von ausweglosen Vertragsbedingungen (ebd., Rn. 85) oder »wenn private Unternehmen in eine staatsähnlich dominante Position rücken […] kann die Grundrechtsbindung

Privater einer Grundrechtsbindung des Staates im Ergebnis vielmehr nahe- oder auch gleichkommen« (ebd., Rn. 88).

Im Urteil zur automatisierten Datenanalyse bei der Polizeiarbeit (BVerfGE 1 BvR 1547/19 und 1 BvR 2634/20 vom 16.2.2023) betont das Gericht u.a., dass aus bereits bestehenden Datensätzen mit automatisierten Datenanalysen neues Wissen, vor allem persönlichkeitsrelevante Informationen, erzeugt werden kann (ebd., Rn. 68). Das Gericht verweist zudem auf die Diskriminierungsrisiken von automatisierten Datenanalysen, die umso weniger hinzunehmen sind, je mehr sich die Wirkungen der Analysen einer unzulässigen Benachteiligung nach Art. 3 (3) GG annähern (ebd., Rn. 76). Darüber hinaus betont es die Nachvollziehbarkeit der Algorithmen für den individuellen Rechtsschutz und die aufsichtliche Kontrolle, um Fehler erkennen und korrigieren zu können. Verlust der staatlichen Kontrolle wird vor allem bei Verwendung von lernfähigen Systemen bzw. KI gesehen, denn diese können im Verlauf des maschinellen Lernprozesses sich von der ursprünglichen menschlichen Programmierung lösen und deren Lernprozesse und Ergebnisse immer schwerer nachzuvollziehen sein (ebd., Rn. 99, mit Verweis auf das EuGH-Urteil Ligue des droits humains, C-817/19).

4. Faktoren der Verletzung der Menschenwürde bei algorithmischen Differenzierungen

4.1 Schwerwiegende und strukturelle Diskriminierung

Algorithmischen Differenzierungen können durch ihre Reichweite, die ganze Populationen umfassen, zu systematischen bzw. strukturellen Diskriminierungen führen. Einige Faktoren können auf eine Verletzung der Menschenwürde durch mögliche systematische algorithmische Diskriminierungen hindeuten, wenn nämlich algorithmische Diskriminierungen nach »Rasse« bzw. Ethnie, Geschlecht, körperlichen und geistigen Behinderungen und den weiteren geschützten Merkmalen des Art. 3 (3) GG erfolgen und wenn algorithmische Differenzierungen in Bereichen eingesetzt werden, bei denen eine starke Abhängigkeit von den Leistungen, Produkten oder Diensten besteht oder besonders Schutzbedürftige betrifft. Dies gilt insbesondere bei Entscheidungen, bei denen es darum geht, ein Leben in Würde zu leben (z.B. bei Empfangenden von Sozialleistungen). Dort kann es bei Diskriminierungen auch zu Formen der Demütigung oder Erniedrigung kommen, dadurch dass eine Behandlung

als Personen mit gleichem moralischem Wert vermissen gelassen wird (siehe z.B. die SyRi und Robo-debt Skandale) (ähnlich Teo 2023, 17).

In dieser Hinsicht sind auch algorithmische Differenzierungen problematisch, die durch negative Rückkopplungsschleifen (feedback loops) strukturelle Ungleichheiten festigen oder ausweiten. Solche negativen Rückkopplungsschleifen können entstehen, wenn Ergebnisse von AES und KI-Systeme, die das Verhalten von Betroffenen vorhersagen, wieder erfasst werden und die Systeme diese Daten unkorrigiert als Datengrundlage für weitere Datenanalysen, Schlussfolgerungen oder Weiterentwicklung bzw. »Lernen« der Algorithmen verwenden. Beispiele finden sich bei Systemen der vorausschauenden Polizeiarbeit (predicitive policing) (Lum und Isaac 2016, FRA 2022). Dies kann ebenso bei generativer KI geschehen, bei der Daten der Kommunikation mit den Nutzenden wieder ausgewertet werden. Darüber hinaus argumentieren Behrendt und Loh (2022), dass negative Rückkopplungsschleifen vor allem bei den ohnehin benachteiligten Bevölkerungsgruppen entstehen, da sie eher dazu tendieren, personenbezogene Daten preiszugeben, was u.a. an der Schwäche des Regulierungsinstruments der informierten Einwilligung (s.u.) und der nur scheinbaren Freiwilligkeit der Preisgabe liegt.

4.2 Generalisierung und fehlende Einzelfallgerechtigkeit

Algorithmische Differenzierung nimmt oft Formen der so genannten statistischen Diskriminierung an und verändert diese (Barocas und Selbst 2016, Binns et al. 2018, Orwat 2019). Statistische Diskriminierung ist eine Art Proxydiskriminierung. Statt mittels einer aufwendigen Einzelfallprüfung das tatsächliche Persönlichkeitsmerkmal bzw. Differenzierungsziel (z.B. das soziale Konstrukt »tatsächliche Fähigkeit, ein Flugzeug zu führen«) zu ermitteln, wird eine vergleichsweise einfach zu erhaltende Proxyinformation (z.B. Alter in Jahren) genutzt. Diese Form der Differenzierung soll ein Informationsdefizit effizient überwinden. Zu einer Diskriminierung kann es kommen, wenn die Proxyinformationen rechtlich geschützte Merkmale sind oder Merkmale enthalten, die eine Korrelation zu geschützten Merkmalen aufweisen (z.B. Hellman 2008, Britz 2008, Schauer 2018). Die Proxyinformationen können aus empirischen Untersuchungen abgeleitet werden oder, im Falle des maschinellen Lernens, auf an Daten trainierten Modellen vorliegen.

Statistische Diskriminierung und Generalisierung (sowohl durch menschliche Entscheidende oder mit Nutzung von Algorithmen) sind jedoch bereits an sich ethisch problematisch, weil Gruppeninformationen auf Individuen über-

tragen werden und so bei Entscheidungen als Quasi-Stereotypen und Vorurteilen bei der Entscheidungsfindung wirken (Gandy Jr. 2010, 34). Prinzipiell ist die Einzelfallgerechtigkeit nicht gewährleistet, da eine Einzelfallprüfung von Personen nicht vorgenommen wird und die individuellen Subjekteigenschaften und individuellen Situationen und Kontexte nicht berücksichtigt werden (Britz 2008). Was in der KI-Forschung und Praxis oft als »Prognose« bezeichnet wird, ist keine Prognose des potenziellen Verhaltens eines Individuums, die aus einer individuellen Prüfung abgeleitet wird. Vielmehr handelt es sich um eine Zuordnung von Personen zu statistisch oder mit maschinellen Lernverfahren gebildeten Kriterien, Kategorien, Scorewerten oder Rangfolgen, von denen bestimmte Ergebnisse in der Zukunft für die zugeordneten Personen erwartet werden.

In vielen Fällen der algorithmischen Differenzierung werden zudem die Kategorien, zu denen Individuen zugeordnet werden, anhand der Daten von Gruppen konstruiert, die nicht die Personen enthalten, über die tatsächlich entschieden wird (Eckhouse et al. 2019, 198f.). Auch sind die konstruierten Kategorien für die betroffenen Personen und Dritte in der Regel nicht nachvollziehbar. Anders als bei der Verwendung von klar kommunizierten Kriterien (z.B. Altersgrenzen in der Verwaltung) entziehen sich solche Entscheidungskriterien und -regeln der Überprüfung und Diskussion durch Wissenschaft und Öffentlichkeit, etwa ob überhaupt ein Kausalzusammenhang zwischen Kriterien und Differenzierungsziel besteht, ob er mit Fakten belegt werden kann oder ob die Verwendung bestimmter Kriterien sozialpolitisch oder moralisch umstritten oder unerwünscht ist.

4.3 Behandlung von Personen als Individuen und mit Achtung

Im Gegensatz zu individualisierten Entscheidungen über Personen werden bei statistischer Diskriminierung und Generalisierung Personen als Informationsobjekte und nicht als Individuen behandelt. Wann es moralisch problematisch ist, Menschen nicht als Individuum zu behandeln, sondern nur als Mitglied einer Gruppe (Stereotypisierung) ist umstritten (z.B. Lippert-Rasmussen 2011, Beeghly 2018).

Häufig wird an den moralischen Überlegungen von Kant angeknüpft, der mit dem Instrumentalisierungsverbot die Achtung des Menschen fordert (nach Hill Jr. 2014, 316f., Dillon 2022, Kapitel 2.2). So ist jeder gehalten, »die Würde der Menschheit an jedem anderen Menschen praktisch anzuerkennen, mithin ruht auf ihm eine Pflicht, die sich auf die jedem anderen Menschen

notwendig zu erzeigende Achtung bezieht.« (Kant 1797/1977, 601). Die Achtung von anderen Personen, die sich Menschen gegenseitig schulden und die Menschen gegenüber anderen gelten machen können, ist die Achtung ihrer Würde (Kant 1797/1977, 600, hier nach Schaber 2016, 256, Ulgen 2017, 2022, 14–15). Eine andere Person (und sich selbst) in seiner Würde achten bedeutet, dass ich andere »jederzeit zugleich als Zweck, niemals bloß als Mittel« (Kant 1786/2021, 429) behandle.

Nach Schaber, der dazu seine Erläuterungen zum falschen Versprechen interpretiert, meint Kant damit auch, dass man eine Person bloß als Mittel behandelt, wenn man sie in einer Weise behandelt, der sie unmöglich zustimmen kann. Das ist der Fall, »wenn sie dazu keinen Grund hat und sie sich nicht rational verhalten würde, wenn sie zustimmen würde.« (Schaber 2016, 256). Eine Person in ihrer Würde achten, bedeutet daher, sie in einer Weise zu behandeln, der sie vernünftigerweise zustimmen kann (ebd.). Bereits Korsgaard (1996), die sich ebenfalls auf Kant bezieht, argumentiert, dass man eine Person dann als bloßes Mittel behandelt, wenn man sie so behandelt, dass die Person der Art der Behandlung nicht zustimmen kann. Dies kann die Person weder bei Zwang noch bei Täuschung, da bei beiden Formen der Behandlung die Person keine Chance bekommen hat, den Zweck auszuwählen. Eine Behandlung ist demnach moralisch schlecht, wenn Personen nicht in der Lage sind zu wählen. Sie schließt daher, dass Zwang und Täuschung, nach der Formel der Menschlichkeit von Kant, die grundlegendsten Formen des Fehlverhaltens gegenüber anderen, die Wurzel allen Übels ist (ebd., S. 137–40). Eine Täuschung ist nach Schaber dann problematisch, wenn sie einen Bereich betrifft, der das Recht der Betroffenen, über wesentliche Teile des eigenen Lebens zu verfügen, beeinträchtigt (Schaber 2013, 134–36).

Ebenso unter Bezugnahme auf Kant entwickelt Ulgen (2022) Anforderungen für die Behandlung von Personen mit Respekt vor der ihnen inhärenten Würde im Hinblick auf KI und AES. Die Würde leitet sich aus ihrer Autonomie und ihren rationalen Fähigkeiten zur Ausübung von Vernunft, Urteilen und Entscheidungen ab. Die menschliche Autonomie ist geschützt, wenn Menschen in der Lage sind, unter dem Einfluss der Vernunft zu handeln, wenn sie die Beweggründe für ihr Handeln erkennen können, oder wenn sie ihre Beweggründe ändern können (ebd., S. 19). Diejenigen KI-Anwendungen und AES, die die Möglichkeiten zur Ausübung von Vernunft, Urteilsvermögen und Wahlmöglichkeiten einschränken, untergraben die Menschenwürde (ebd., S. 27). Neben der Relevanz für technologisch implementierte soziale Regeln

(s.u.) ist dies bereits für die algorithmenbasierte Entscheidungsfindung relevant.

Diese Argumentationen zeigen, wie bedeutend die Vorbedingungen für die Art und Weise der Behandlung sowie funktionsfähige Kompensationsmechanismen sind, um zu verhindern, dass Individuen als bloßes Objekt behandelt werden. Insbesondere gehören dazu die Möglichkeiten der Zustimmung zur Behandlung und der Einwirkung auf die Behandlung. Auch die Sicherung von Wahlmöglichkeiten und die Voraussetzung, über die Wahlmöglichkeiten so informiert zu werden, dass Personen selbstbestimmt handeln können, gehören zum Schutz der Menschenwürde.

Philosophische Erklärungsansätze dafür, wann eine Differenzierung moralisch falsch ist, ziehen vergleichbare Schlüsse. Zu ihnen zählen an Menschenwürde und Missachtung orientierte Ansätze der Diskriminierungstheorie (Khaitan 2015, 6–8), auch wenn nicht immer der Begriff »Würde« verwendet wird. Sie sehen eine Differenzierung als falsch an, wenn die diskriminierende Person den moralischen Wert der diskriminierten Person falsch, vor allem als niedriger, einschätzt oder wenn die diskriminierende Person eine falsche Einschätzung zum Ausdruck bringt, also so handelt, als ob die diskriminierte Person einen geringeren moralischen Wert hat (Thomsen 2017).

Baer (2009) sieht Gleichheitsbedürfnisse nicht allein durch Gleichheitsvorstellungen, sondern besser auch durch Bezugnahme auf Freiheit und Würde befriedigt, da sie als Schutzschild gegen kollektivistische Stereotypisierung dienen. Würde ist das Versprechen der Anerkennung unterschiedlicher Selbstwahrnehmungen, die alle den gleichen Respekt verdienen. Aus dem Zusammenwirken von Gleichheit, Freiheit und Würde erwächst nicht nur das Verständnis, dass Menschenwürde für alle Menschen gleich ist, unabhängig von Status, Klasse oder ähnlichem. Sondern Freiheit stellt auch sicher, dass jeder Einzelne sein eigenes Selbstverständnis definiert, anstatt es von Autoritäten bestimmen zu lassen (ebd., 460). Ebenso kann daraus ein Recht auf Entscheidung unter den Bedingungen der Chancengleichheit frei von Unterdrückung und Unterordnung (Gleichheit) konzipiert werden, unter Achtung und Anerkennung aller Beteiligten (Würde) (ebd., 466).

Hellman (2008) sieht das moralisch Falsche einer Diskriminierung darin, dass sie eine Person erniedrigt. Erniedrigende Regeln oder Praktiken drücken eine Missachtung der moralischen Gleichheit der von Diskriminierung betroffenen Personen aus. Eidelson (2013) fasst das Falsche an Diskriminierung als Fehler, eine Person korrekt als Individuum zu behandeln, auf. Der Fehler liegt darin, die Person nicht als (teilweise) Ergebnis ihrer vergangenen Bemühun-

gen der Selbstentfaltung (self-creation) zu sehen und als autonom Handelnde, deren künftige Entscheidungen sie selbst treffen kann (ebd., 227). So beantwortet er das Problem, dass Personen bei statistischer Diskriminierung und Generalisierung nicht als Individuum behandelt werden, mit der Forderung, dass Personen dann und nur dann als Individuum behandelt werden, wenn (1.) die differenzierende Person X den Nachweisen (evidence), wie die betroffene Person Y ihre Autonomie bei der Gestaltung ihres Lebens ausgeübt hat, ein angemessenes Gewicht beimisst, sofern diese Hinweise in angemessener Weise verfügbar und für die vorliegende Entscheidung relevant sind. (2.) Zusätzlich dürfen die Beurteilungen von X, wenn sie die Auswahlentscheidungen der Person Y betreffen, nicht in einer Art erfolgen, die die Fähigkeit von Y, diese Auswahlentscheidungen als autonom handelnde Person zu treffen, herabsetzen (Eidelson 2015, 144).

Zwar bleiben noch Fragen nach dem Ausmaß, Typen von oder Verpflichtungen zu angemessenen und relevanten Nachweisen offen, doch es kann abgeleitet werden, dass der Gegenstand der Informationen und Entscheidungen die selbstbestimmte Persönlichkeitsentfaltung der Betroffenen, ihre Möglichkeiten der Selbstwahrnehmung, Selbstbestimmung und Selbstdarstellung sein sollte. Allerdings muss ein Dilemma vermieden werden: Sollen möglichst viele personenbezogene Daten über die Selbstbestimmung erfasst werden, um das Problem der Generalisierung zu lösen und Personen als Individuum besser zu achten, kann dies nur mit der Kontrolle der dazu zu erstellenden Daten und Profile durch die Betroffenen selbst geschehen, um eine Verletzung des Rechts auf informationelle Selbstbestimmung zu vermeiden. Statt automatisierte Datenerfassungen und -analysen können auch menschliche Entscheidende erforderlich werden, um kontext-, situations- und personenbezogene Entscheidungen treffen zu können, die einen hohen Grad an situativem Abwägen mit Ermessensspielräumen erfordern.

4.4 Automatisierte Entscheidungen

Automatische Entscheidungen auf Basis von Generalisierungen sind nach Citron und Kaminski moralisch problematisch, da außer den Generalisierungsinformationen keine anderen Informationen über die Betroffenen verarbeitet werden. Werden Individuen bloß algorithmisch gebildeten Kategorien, Scorewerten oder Rangfolgen zugeordnet, werden sie nicht mehr als Individuen behandelt. Wenn es bei algorithmischen Entscheidungen den Individuen nicht mehr möglich ist, ihre Individualität zu verdeutlichen, dann verletzt das ih-

re Würde und verdinglicht sie anhand weniger Merkmale, anstatt sie als ganze Personen zu behandeln. Sowohl die Ausübung menschlichen Ermessens als auch individuelle Verfahrensrechte (des Einspruchs, der Korrektur etc.) sind nicht nur notwendig, um Fehler zu vermeiden, sondern auch, damit die Individualität angemessen anerkannt und respektiert werden kann (Citron 2008, 1304, Kaminski 2019, 1541–45). Des Weiteren sind menschliche Ermessensentscheidungen notwendig, wenn in Entscheidungen auch mildernde Umstände einbezogen werden müssen, die der Algorithmus nicht berücksichtigen kann, ebenso wenn unbestimmte Begriffe in den Entscheidungsregeln bestehen, die vom menschlichen Entscheidenden Abwägungen zwischen gegenläufigen Interessen erfordern (Citron 2008, 1304).

Eine der Beweggründe für die Regulierung der automatisierten Entscheidungsfindung ist der Schutz der Menschenwürde. Dies bezieht sich auf Artikel 22 der Allgemeinen Datenschutzgrundverordnung (Verordnung 2016/679 (DSGVO)) und den Vorgänger, Artikel 15 der Datenschutzrichtlinie (95/46/EG, 1995 (DSRL)). Nach Dammann und Simitis (1997, 218f.) sollte mit dem Verbot automatisierter Entscheidungen (Artikel 15 DSRL) verhindert werden, dass Betroffene bei Persönlichkeitsbeurteilungen nur computergestützt und auf der Grundlage gespeicherter Daten behandelt werden. Dies ignoriere die Individualität der Person und werte die Person zu einem bloßen Objekt von Computeroperationen ab (ähnlich zur DSGVO Martini 2021, Rn. 8, Scholz 2019, Rn. 3, ähnlich Jones 2017, Kaminski 2019).

Nach Martini und Nink wird die Subjektqualität eines Menschen allerdings nicht notwendig dadurch missachtet, dass allein personenbezogene Daten das Objekt einer algorithmischen Analyse sind. Die Subjektqualität wird bei automatisierten (Verwaltungs-)Entscheidungen erst dann tangiert, wenn algorithmische Verfahren den Betroffenen nachteilige Folgen aufbürdet, »ohne ihm die Chance zu eröffnen, sich gegen die Entscheidung in angemessener Weise zur Wehr setzen zu können.« (Martini und Nink 2017, 7). Zur Wahrung der informationellen Selbstbestimmung setzt man in der Rechtspraxis auf das Verfahren, (1) über die automatisierte Entscheidung zu informieren, (2) auf Anfrage die wesentlichen Entscheidungsgründe mitzuteilen und zu erläutern, (3) den eigenen Standpunkt geltend machen zu können, um erforderlichenfalls eine Überprüfung und Neubewertung zu erreichen (ebd., S. 7).

Allerdings lassen einige Defizite der Regulierung von automatisierten Entscheidungen Zweifel aufkommen, ob sie dem Schutz der Menschenwürde noch dienen kann. Das so genannte »Verbot« ist mit umfangreichen Ausnah-

men versehen, insbesondere wenn ein AES zum Abschluss oder der Erfüllung eines Vertrags dient, durch Rechtsvorschriften Zulässigkeit verlangt oder eine ausdrückliche Einwilligung vorliegt. Die Regulierung sieht zwar vor, dass der Betreibenden einer automatisierten Entscheidung über das Bestehen einer automatischen Entscheidung und die so genannte involvierte Logik informieren muss, aber es ist noch unklar, welche Inhalte diese Informationspflicht hat, z.B. ob und wie über Entscheidungskriterien oder mögliche Diskriminierungsrisiken informiert werden muss (Orwat 2019, 114–23 m.w.N.). Zudem wird das »Verbot« oft nur als Eingriffsrecht der Betroffenen in begründeten Einzelfällen interpretiert (Martini und Nink 2017, 4). Dabei müssen die Betroffenen zunächst Kenntnis von dem AES und deren Auswirkungen haben und eine Begründung des Verlangens nach einem Eingreifen eines Menschen und nach Erklärung der involvierten Logik erbringen. Da dies sehr aufwendig sein kann, können Abschreckungseffekte (chilling effects) entstehen, wenn einzelne Personen es als unzumutbare hohe Hürden wahrnehmen, die Regelung in Anspruch zu nehmen.

4.5 Entstehung neuen Wissens sowie umfassender und aussagekräftiger Personen- und Gruppenprofile

Die Möglichkeiten der Datenaggregation, der Wiederverwendung von Daten, der Datenkombination und daraus abgeleitete Schlussfolgerungen, der De-Anonymisierung und der Re-Identifizierung von Personen, der Kategorisierung, Einstufung, Beurteilung und des Individual- oder Gruppen-Profiling von Personen sind mit KI stark angewachsen (z.B. Yeung 2019, FRA 2020, Smuha 2021). Einige KI-Systeme wurden dafür entwickelt, automatisierte Rückschlüsse auf die Identität, persönlichkeitskonstituierenden Merkmale und andere sensible Sachverhalte, wie Emotionen, Charaktereigenschaften, psychische Zustände oder politische Orientierungen zu ziehen (Beispiele in Kosinski 2021, Matz et al. 2023). Die KI-basierten biometrischen und psychometrischen Auswertungen (z.B. emotional AI) können der gezielten Ansprache (z.B. im Marketing), der Risikobeurteilung (z.B. bei der Bewerberselektion, Berechnung der Wahrscheinlichkeit des Studienabbruchs oder Kreditausfalls) und Verhaltenssteuerung dienen (kritisch z.B. Valcke, Clifford, und Dessers 2021). Oft basieren die Systeme auf einer Reduktion der Persönlichkeit auf quantifizierbare Messgrößen und Klassen, die versuchen, die für ein Differenzierungsziel relevanten Persönlichkeitseigenschaften abzubilden. Kritisch wird dazu die Standardisierung von Persönlichkeit (Köchling et al.

2021) oder die pseudowissenschaftliche Herangehensweise (Sloane, Moss, und Chowdhury 2022) angeführt.

Auch wenn Umfang und Arten der Anwendung solche Systeme in der Praxis noch wenig bekannt sind, verdeutlicht dies, dass mit KI aussagekräftige Personenprofile bzw. »Rundum«-Profile erzeugt und verwendet werden können, die geeignet sind, ein (nahezu) vollständiges Fremdbild einer Person überzustülpen, dies mit persönlichkeitskonstituierenden Merkmalen und dies auch ohne dass eine valide Zustimmung durch die Betroffenen vorliegt.

Insgesamt kommt es zu einer weiteren Ablösung der Datenrepräsentation durch die Betreibenden von den Möglichkeiten der Kontrolle der Selbstdarstellung durch die Betroffenen (vgl. auch Teo 2023). Die Identität der Betroffenen ist dann nur noch fremdbestimmt, auch wie sie sich begreifen (als normal, gesund, regelkonform etc.). Die Fähigkeiten, sich selbstbestimmt zu entfalten, sind dann eingeschränkt oder sogar beseitigt. Es liegt nach den oben skizzierten Maßstäben eine Verletzung der informationellen Selbstbestimmung und der Menschenwürde vor. Auch können derartige KI-Anwendungen Abschreckungseffekte und damit Selbsteinschränkungen der freien Persönlichkeitsentfaltung als Form der Würdeverletzung hervorrufen (FRA 2019, 20).

4.6 Strukturelle Überlegenheit

Im Verhältnis Staat gegenüber Betroffenen (z.B. Bürgern oder Migranten) muss man grundsätzlich von der strukturellen Überlegenheit des Staates ausgehen, denn es liegen üblicherweise Situationen mit Gewaltmonopol, fehlende Ausweichmöglichkeiten, Ausgeliefertsein, Nichtverhandelbarkeit und vollständiger, einseitig festgelegter Verbindlichkeit der Regeln vor. Eine Reihe von Faktoren können auch in privaten Verhältnissen die strukturelle Überlegenheit von Anwendenden der KI-Systeme (z.B. Anbietende, Arbeitgebende, Banken) gegenüber den Betroffenen (z.B. Kunden, Bewerbende, Kreditsuchenden) erhöhen.

Erstens werden sowohl im staatlichen als auch im privaten Bereich zunehmend gesellschaftliche Regeln in Software bzw. Algorithmen gefasst. Für die technische Implementierung müssen die Regeln in Programmiersprache umgesetzt werden oder werden durch maschinelles Lernen erzeugt. Bei diesen Formen der Präzisierung gehen aber auch Auslegungs- und Ermessensspielräume verloren, die oft notwendig sind, damit gesellschaftliche Regeln auf viele Situationen, die teils nicht vorhersehbar sind, angewandt werden können. Werden Regeln vollständig automatisiert durchgesetzt, wie z.B. bei

vollautomatisierten Entscheidungen, wird ein Abweichen von den Regeln technisch verhindert. Aber auch bei algorithmischen Auswahlarchitekturen (choice architectures oder nudging) wird der Raum der Auswahlmöglichkeiten technisch vorgegeben und oft verengt. Die Regeln werden dann zumeist einseitig von den Entwickelnden und Anwendenden vorgegeben, Verhandlungs-, Einwirkungs-, Korrekturmöglichkeiten der Betroffenen reduziert oder beseitigt, wodurch sich die strukturelle Überlegenheit erhöhen kann. Dadurch verringern sich auch die Handlungsmöglichkeiten, die Autonomie und die Chancen der Selbstdurchsetzung von Autonomie durch die Betroffenen (Ulgen 2022, Teo 2023, 27–31, Deutscher Ethikrat 2023, 120–37).

Zweitens, wenn die privaten Anwendenden von KI-Systemen gleichzeitig auch diejenigen sind, die Plattformen betreiben (die teilweise auch die Entwickelnden von KI-Systemen sind), können starke Netzwerkeffekte der Plattformen zu verminderten Ausweichmöglichkeiten und stärkeren Abhängigkeiten führen.

Drittens sind, wie eben gezeigt, einige KI-Systeme in der Lage, sensible Persönlichkeitsmerkmale, wie psychische Zustände, Charaktereigenschaften, Emotionen, auch aus scheinbar »belanglosen« Daten, wie der Kommunikation in sozialen Netzen, zu ermitteln. Dadurch kann die Angewiesenheit auf ein Produkt, Dienst oder Position besser ermittelt und ausgenutzt werden (z.B. Härtel 2019), ebenso menschliche Schwächen, insbesondere wenn die Systeme für »Dark Pattern« oder anderen Formen der Manipulation verwendet werden (z.B. Ulgen 2022, 22–24).

Entwicklungen, die die strukturelle Überlegenheit der Anwendenden von KI-Systemen gegenüber den Betroffenen erhöhen, können tendenziell die Menschenwürde und die freie Persönlichkeitsentfaltung gefährden, denn den Betroffenen werden die Möglichkeit der selbstbestimmten Lebensgestaltung eingeschränkt. Dies kann in privaten Verhältnissen auf dem Wege der Störung der Vertragsparität und damit der Verringerung der Möglichkeiten, dass Betroffene ihre Autonomie selbst durchsetzen können, geschehen. Denn auch wenn die Vertragsfreiheit gilt, d.h. dass jeder die Freiheit hat, zu bestimmen, mit wem und unter welchen Bedingungen Verträge abgeschlossen werden, so muss gesichert sein, dass dabei »auch die Bedingungen freier Selbstbestimmung tatsächlich gegeben sind.« (BVerfGE 82, 242, S. 255).

4.7 Fehlende Möglichkeit der validen Zustimmung

Das Instrument der Zustimmung kann moralisch unzulässige Behandlungen in zulässige transformieren. Allerdings kann sie diese moralische Transformationsleistung nur bei Einhaltung bestimmter Voraussetzungen erreichen. Dazu gehört, dass die Betroffenen freiwillig zustimmen und dazu Wahlmöglichkeiten haben, dass sie ausreichend informiert sind, also den Umfang der Datenverarbeitungen und die Konsequenzen daraus verstehen, und dass sie die notwendigen Entscheidungsbefähigungen haben (Bullock 2018). Zum anderen wird aus philosophischer Sicht auch bezweifelt, dass Zustimmung eine Behandlung als bloßes Objekt in eine moralisch zulässige Behandlung transformieren kann, wenn bereits die Behandlung die Pflicht, andere Menschen mit Achtung zu behandeln, verletzt (Fahmy 2023). Das dürfte etwa bei schwerwiegenden algorithmischen Diskriminierungen oder bei Differenzierungen, die auf Profilen mit (nahezu) vollständiger Erfassung und Fremdbestimmung der Persönlichkeit basieren, der Fall sein.

In der Datenschutzpraxis sind die Probleme der informierten Einwilligung seit langem bekannt. Die Wirksamkeit und Aussagekraft wird zunehmend eingeschränkt durch nicht verhandelbare, lange, unverständliche und in juristischer Sprache formulierte Datenschutzerklärungen, die zunehmende Erhebung von Daten auf der Grundlage sogenannter berechtigter Interessen ohne das einer Einwilligung erforderlich ist (Artikel 6 (1) f DSGVO), starke technisch-ökonomische Netzwerkeffekte und dadurch Bindungen von Kunden und Nutzern an Systeme oder Plattformen (abnehmende Freiwilligkeit) sowie durch Schnittstellendesigns, die zu einer Einwilligung in Datenerhebungen verleiten. Den Betroffenen fehlt oft die Kenntnis über die Notwendigkeit und die rechtlichen Möglichkeiten der informierten Einwilligung selbst. Zudem können sie kaum abschätzen, welche tatsächlichen Folgen die Einwilligung im Hinblick auf potenziell nachteilige, teilweise zeitlich weit entfernte Behandlungen, die auch auf Basis schwer nachvollziehbarer Datenakkumulation oder nicht mehr abschätzbarer Weiterverarbeitungen und Datenweitergaben entstehen können, hat. Darüber hinaus können bei komplexen KI-Algorithmen die Entscheidungskriterien unverständlich oder unbekannt sein, insbesondere wenn es sich um selbstlernende bzw. adaptive Systeme handelt. Ebenso kann mit KI-basierten Schlussfolgerungen neues Wissen aus vorhandenen personenbezogenen Daten generiert werden, auch aus anonymisierten Daten und auch für Personen oder Gruppen, die nicht an der ursprünglichen Einwilligung beteiligt waren (s.o.). Es ist dann davon auszugehen, dass die

Betroffenen nicht mehr hinreichend erkennen können, in was sie einwilligen (z.B. Orwat 2019, 106f. m.w.N.). Auf Grund dieser Faktoren wird die informierte Einwilligung als Legitimation der Behandlung von Menschen als bloße Objekte und als Instrument der Selbstbestimmung zunehmend unbrauchbar.

Zarsky konkretisiert noch detaillierter, dass zum Schutz der Menschenwürde ein Verständnis der inneren Abläufe von automatisierten Datenanalysen vorliegen müsse, denn ohne dieses Verständnis können die Ergebnisse immer noch willkürlich und falsch erscheinen. Das Problem ist, dass mit den bestehenden (rechtlichen) Transparenzanforderungen maximal lediglich Informationen über die Korrelationen bzw. Klassifikationen, in denen eine Person eventuell fallen könnte, geliefert werden. Stattdessen muss der automatisierte Prognoseprozess auch interpretierbar sein, d.h. der Auswahlprozess muss erklärbar sein. Daher erfordere der Schutz der Menschenwürde sogar, dass für die Betroffenen kausale Zusammenhänge und nicht bloß Korrelationen feststellbar sein müssten, bevor Schlussfolgerungen und Maßnahmen getroffen werden (Zarsky 2013).

5. Zusammenfassung und Schlussfolgerung

Die Menschenwürdeperspektive ermöglicht es zu bestimmen, wie Menschen oder Maschinen andere Menschen behandeln sollten. Diese Perspektive kann die Arbeiten zur relativen Fairness von Algorithmen und der Verminderung von Diskriminierungsrisiken ergänzen. Die Menschenwürdeperspektive geht über das Bemühen des »debiasing« von Datensätzen und Algorithmen hinaus und fragt, wie algorithmenbasierte Entscheidungsprozesse gestaltet sein sollten und welche Informationsgrundlagen dafür zur Verfügung stehen sollten. Sie weist darauf hin, dass selbst der Idealfall »genauer« Profile als Grundlage algorithmenbasierter Entscheidungen problematisch ist, wenn übermächtige Fremdbilder die informationelle Selbstbestimmung der Betroffenen unterdrücken. Sie liefert auch Begründungen dafür, in welchen Situationen KI und AES nicht eingesetzt werden sollten, weil eine Einschränkung oder Verletzung der Menschenwürde nicht ausgeschlossen werden kann. So sind die Verbote bestimmter KI-Anwendungen im Entwurf der KI-Verordnung auch mit Verweis auf den Schutz der Menschenwürde begründet worden (z.B. in Erwägung 15 und 17). Das Verständnis von Diskriminierung auch als Verletzung der Würde und der moralischen Gleichwertigkeit der Betroffenen ergänzt das Verständnis von Diskriminierung als Schädigungen von Gerechtigkeitsvor-

stellungen oder von sozialpolitischen Zielen. Es gibt Aufschluss darüber, was es bedeutet, Menschen als Individuen und mit Respekt zu behandeln.

Durch den Einsatz von KI und AES kann es zur Verletzung der Menschenwürde kommen. Wie gezeigt, kann dies mit der Überlagerung von problematischen Faktoren bei der Anwendung von KI und AES und die zunehmende Ungeeignetheit von rechtlich etablierten Kompensationsmechanismen zur Abmilderung der Behandlung als bloßem Objekt geschehen. Diese Faktoren umfassen (1.) die Generalisierung und Missachtung der Persönlichkeit in Entscheidungen der Ungleichbehandlung, (2.) die Reichweite von Systemen mit Restrisiken der systematischen und strukturellen Diskriminierung, einschließlich dem Umstand, einige Personen einem höheren Diskriminierungsrisiko auszusetzen und sie wie Personen mit geringerem moralischen Wert zu behandeln, (3.) die immer unzureichender werdende informierte Einwilligung, die sich bei KI-Systemen, deren Entscheidungskriterien und Auswirkungen nicht mehr nachvollziehbar sind, besonders drastisch auswirkt und Betroffene nicht mehr auf die Ergebnisse einwirken können, (4.) die unzureichende Klärung der Regulierung von automatisierten Entscheidungen, der Rolle involvierter menschlicher Entscheidender sowie der informierten Einwilligung dabei, (5.) der Verlust der Kontrolle über die Erzeugung und Verwendung von Persönlichkeitsbildern durch die Betroffenen sowie (6.) die steigende strukturelle Überlegenheit des Staates oder privater Unternehmen durch die zunehmende technische Durchsetzung von gesellschaftlichen Regeln, die Marktkonzentration, die besonderen Fähigkeiten von KI-Systemen Abhängigkeiten und andere menschliche Schwächen zu erkennen und sie auszunutzen und dadurch Situationen mit eingeschränkten Handlungs- und Einwirkungsmöglichkeiten, Unausweichlichkeit und starken Abhängigkeiten. Als Folge kann in den Situationen, in denen Faktoren allein oder zusammenwirken, eine Garantie des Schutzes der Menschenwürde nicht mehr vorliegen. Dies wiegt besonders schwer, wenn es sich um Differenzierungen von Produkten, Diensten, Positionen handelt, die für die selbstbestimmte Lebensgestaltung und Identitätsbildung oder für ein menschenwürdiges Dasein von Menschen mit besonderen Bedürfnissen und Vulnerabilitäten essentiell sind.

Es besteht eine dringende Notwendigkeit, weiter zu klären, wann die Menschenwürde und Persönlichkeitsentfaltung konkret eingeschränkt oder verletzt ist und wie sie zu schützen sind, insbesondere (1.) welche Personen- und Gruppenprofile so umfassend oder so persönlichkeitskonstituierend sind, dass man das Persönlichkeitsbild als fremdbestimmt und den Kernbereich

der privaten Lebensgestaltung als ausgehöhlt bezeichnen muss, (2.) unter welchen Bedingungen schwerwiegende, systematische oder strukturelle Diskriminierungen mit Verwendung von Algorithmen vorliegen, (3.) wie weit und in welcher Form die Persönlichkeit der Betroffenen bei algorithmenbasierten Entscheidungen zu respektieren ist und welche Form von Rechtfertigung für Entscheidungen Betroffene erhalten müssen, (4.) welche Einwirkungsmöglichkeiten auf Entscheidungen und auf ihr Persönlichkeitsbild die Betroffenen haben müssen und wie kommunikative Prozesse dazu aussehen sollten sowie (5.) wie nicht nur die Menschenwürde und Persönlichkeitsentfaltung der direkt Betroffenen (wieder) gestärkt werden kann, sondern auch der Schutz von mitbetroffenen Dritten, die nicht wissen, dass sie betroffen sind.

Anmerkung und Danksagung

Eine ähnliche Version dieses Beitrags soll parallel in englischer Sprache erscheinen. Für wertvolle Hinweise bei der Erarbeitung dieses Beitrags möchte ich meinen Kollegen Reinhard Heil und Philipp Frey danken.

Literatur

Baer, Susanne. 2009. »Dignity, liberty, equality: A fundamental rights triangle of constitutionalism.« *University of Toronto Law Journal* 59 (4): 417–68.

Barocas, Solon and Andrew D. Selbst. 2016. »Big Data's Disparate Impact.« *California Law Review* 104 (3): 671–732.

Beeghly, Erin. 2018. »Failing to treat persons as individuals.« *Ergo: An Open Access Journal of Philosophy* 5 (26): 687–711.

Behrendt, Hauke and Wulf Loh. 2022. »Informed consent and algorithmic discrimination – is giving away your data the new vulnerable?« *Review of Social Economy* 80 (1): 58–84.

Bender, Emily M., Timnit Gebru, Angelica McMillan-Major and Shmargaret Shmitchell. 2021. »On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?« *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*.

Binns, Reuben, Max Van Kleek, Michael Veale, Ulrik Lyngs, Jun Zhao and Nigel Shadbolt. 2018. »›It's Reducing a Human Being to a Percentage‹: Percep-

tions of Justice in Algorithmic Decisions.« *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*.

Britz, Gabriele. 2007. *Freie Entfaltung durch Selbstdarstellung. Eine Rekonstruktion des allgemeinen Persönlichkeitsrechts aus Art. 2 I GG*. Tübingen: Mohr Siebeck.

Britz, Gabriele. 2008. *Einzelfallgerechtigkeit versus Generalisierung. Verfassungsrechtliche Grenzen statistischer Diskriminierung*. Tübingen: Mohr Siebeck.

Britz, Gabriele. 2010. »Informationelle Selbstbestimmung zwischen rechtswissenschaftlicher Grundsatzkritik und Beharren des Bundesverfassungsgerichts.« In *Offene Rechtswissenschaft*, edited by Wolfgang Hoffmann-Riem, 561–96. Tübingen: Mohr Siebeck.

Bullock, Emma C. 2018. »Valid consent.« In *The Routledge Handbook of the Ethics of Consent*, edited by Peter Schaber and Andreas Müller, 85–94. Routledge.

Buolamwini, Joy and Timnit Gebru. 2018. »Gender shades: Intersectional accuracy disparities in commercial gender classification.« *Conference on Fairness, Accountability and Transparency*.

Citron, Danielle K. 2008. »Technological Due Process.« *Washington University Law Review* 85 (6): 1249–313.

Dammann, Ulrich and Spiros Simitis. 1997. *EG-Datenschutzrichtlinie: Kommentar*. Baden-Baden: Nomos.

Deutscher Ethikrat. 2023. *Mensch und Maschine – Herausforderungen durch Künstliche Intelligenz*. Deutscher Ethikrat. Berlin.

Dillon, Robin S. 2022. »Respect.« In *The Stanford Encyclopedia of Philosophy* (Fall 2022 Edition), edited by Edward N. Zalta and Uri Nodelman. Metaphysics Research Lab, Stanford University.

Eckhouse, Laurel, Kristian Lum, Cynthia Conti-Cook and Julie Ciccolini. 2019. »Layers of Bias: A Unified Approach for Understanding Problems With Risk Assessment.« *Criminal Justice and Behavior* 46 (2): 185–209.

Eidelson, Benjamin. 2013. »Treating People as Individuals.« In *Philosophical Foundations of Discrimination Law*, edited by Deborah Hellman and Sophia Moreau, 203–27. Oxford: Oxford University Press.

Eidelson, Benjamin. 2015. *Discrimination and Disrespect*. Oxford: Oxford University Press.

Fahmy, Melissa S. 2023. »Never Merely as a Means: Rethinking the Role and Relevance of Consent.« *Kantian Review* 28 (1): 41–62.

FRA. 2019. *Facial recognition technology: fundamental rights considerations in the context of law enforcement*. European Union Agency for Fundamental Rights (FRA). Luxembourg: Publications Office of the European Union.

FRA. 2020. *Getting the Future Right – Artificial Intelligence and Fundamental Rights*. European Union Agency for Fundamental Rights (FRA). Luxembourg: Publications Office of the European Union.

FRA. 2022. *Bias in Algorithms – Artificial Intelligence and Discrimination*. European Union Agency for Fundamental Rights (FRA). Luxembourg: Publications Office of the European Union.

Gandy Jr., Oscar H. 2010. »Engaging rational discrimination: exploring reasons for placing regulatory constraints on decision support systems.« *Ethics and Information Technology* 12 (1): 1–14.

Hacker, Philipp. 2018. »Teaching Fairness to Artificial Intelligence: Existing and Novel Strategies Against Algorithmic Discrimination Under EU Law.« *Common Market Law Review* 55 (4): 1143–85.

Härtel, Ines. 2019. »Digitalisierung im Lichte des Verfassungsrechts – Algorithmen, Predictive Policing, autonomes Fahren.« *Landes- und Kommunalverwaltung* 29 (2): 49–60.

Hellman, Deborah. 2008. *When is Discrimination Wrong?* Cambridge, London: Harvard University Press.

Herdegen, Matthias. 2022. »Art. 1 Abs. GG.« In *Grundgesetz Kommentar*, edited by Theodor Maunz and Günther Dürig. München: Beck.

Hill Jr., Thomas E. 2014. »In Defence of Human Dignity: Comments on Kant and Rosen.« In *Understanding Human Dignity*, edited by Christopher McCrudden, 313–25. Oxford: Oxford University Press.

Hillgruber, Christian. 2023. »GG Art. 1 Schutz der Menschenwürde.« In *Beck Online-Kommentar Grundgesetz*, edited by Volker Epping and Christian Hillgruber. München.

Höfling, Wolfram. 2021. »Art. 1 GG Schutz der Menschenwürde, Menschenrechte, Grundrechtsbindung.« In *Grundgesetz: Kommentar*, edited by Michael Sachs, 70–102. München: Beck.

Hong, Mathias. 2019. *Der Menschenwürdegehalt der Grundrechte. Grundfragen, Entstehung und Rechtsprechung*. Tübingen: Mohr Siebeck.

Hong, Mathias. 2022. »Grundwerte des Antidiskriminierungsrechts: Würde, Freiheit, Gleichheit und Demokratie.« In *Handbuch Antidiskriminierungsrecht. Strukturen, Rechtsfiguren und Konzepte*, edited by A.K. Mangold and M. Payandeh, 67–123. Tübingen: Mohr Siebeck.

Jones, Meg L. 2017. »The right to a human in the loop: Political constructions of computer automation and personhood.« *Social Studies of Science* 47 (2): 216–39.

Kaminski, Margot E. 2019. »Binary Governance: Lessons from the GDPR's Approach to Algorithmic Accountability.« *Southern California Law Review* 92 (6): 1529–616.

Kant, Immanuel. 1786/2021. *Grundlegung der Metaphysik der Sitten*, edited by Theodor Valentiner, Seitenzahl nach Akademieausgabe Band IV. Stuttgart: Reclam.

Kant, Immanuel. 1797/1977. *Die Metaphysik der Sitten*, edited by Wilhelm Weischedel. Frankfurt a.M.: Suhrkamp.

Khaitan, Tarunabh. 2015. *A Theory of Discrimination Law*. Oxford: Oxford University Press.

Köchling, Alina, Shirin Riazy, Marius C. Wehner and Katharina Simbeck. 2021. »Highly Accurate, But Still Discriminatory.« *Business & Information Systems Engineering* 63 (1): 39–54.

Korsgaard, Christine M. 1996. *Creating the Kingdom of Ends*. Cambridge: Cambridge University Press.

Kosinski, Michal. 2021. »Facial recognition technology can expose political orientation from naturalistic facial images.« *Scientific Reports* 11 (1): Article 100 (7 pages).

Lehner, Roman. 2013. *Zivilrechtlicher Diskriminierungsschutz und Grundrechte. Auch eine grundrechtliche Betrachtung des 3. und 4. Abschnittes des Allgemeinen Gleichbehandlungsgesetzes (§§19-23 AGG)*. Tübingen: Mohr Siebeck.

Lippert-Rasmussen, Kasper. 2011. «We are all Different«: Statistical Discrimination and the Right to be Treated as an Individual.« *The Journal of Ethics* 15 (1): 47–59.

Lum, Kristian and William Isaac. 2016. »To predict and serve?« *Significance* 13 (5): 14–19.

Mahlmann, Matthias. 2008. *Elemente einer ethischen Grundrechtstheorie*. Baden-Baden: Nomos.

Mahlmann, Matthias. 2012. »Human Dignity and Autonomy in Modern Constitutional Orders.« In *The Oxford Handbook of Comparative Constitutional Law*, edited by Michael Rosenfeld and András Sajó, 1–26. Oxford: Oxford University Press.

Martini, Mario. 2021. »DS-GVO Art. 22 Automatisierte Entscheidungen im Einzelfall einschließlich Profiling.« In *Datenschutz-Grundverordnung Bundesdatenschutzgesetz DS-GVO BDSG, Kommentar*, 3. Auflage, edited by Boris P. Paal and Daniel A. Pauly. München: C.H. Beck.

Martini, Mario and David Nink. 2017. »Wenn Maschinen entscheiden… – vollautomatisierte Verwaltungsverfahren und der Persönlichkeitsschutz.« *Neue Zeitschrift für Verwaltungsrecht – Extra* 36 (10): 1–14.

Matz, Sandra C., Christina S. Bukow, Heinrich Peters, Christine Deacons, Alice Dinu and Clemens Stachl. 2023. »sing machine learning to predict student retention from socio-demographic characteristics and app-based engagement metrics.« *Scientific Reports* 13 (1): Article 5705 (16 pages).

McCrudden, Christopher. 2008. »Human dignity and judicial interpretation of human rights.« *European Journal of International Law* 19 (4): 655–724.

Mehrabi, Ninareh, Fred Morstatter, Nripsuta Saxena, Kristina Lerman and Aram Galstyan. 2021. »A survey on bias and fairness in machine learning.« *ACM Computing Surveys* 54 (6): 1–35.

Obermeyer, Ziad, Brian Powers, Christine Vogeli and Sendhil Mullainathan. 2019. »Dissecting racial bias in an algorithm used to manage the health of populations.« Science 366 (6464): 447–53.

Orwat, Carsten. 2019. *Diskriminierungsrisiken durch Verwendung von Algorithmen*. Studie erstellt mit einer Zuwendung der Antidiskriminierungsstelle des Bundes. Berlin: Nomos.

Pessach, Dana and Erez Shmueli. 2022. »A review on fairness in machine learning.« *ACM Computing Surveys (CSUR)* 55 (3): 1–44.

Schaber, Peter. 2013. *Instrumentalisierung und Menschenwürde*. 2 ed. Münster: Mentis.

Schaber, Peter. 2016. »Menschenwürde.« In *Handbuch Gerechtigkeit*, edited by Anna Goppel, Corinna Mieth and Christian Neuhäuser, 256–62. Stuttgart: J.B. Metzler.

Schauer, Frederick. 2018. »Statistical (and non-statistical) discrimination.« In *The Routledge Handbook of the Ethics of Discrimination*, edited by Kasper Lippert-Rasmussen, 42–53. London: Routlegde.

Scholz, Philip. 2019. »DSGVO Art. 22 Automatisierte Entscheidungen im Einzelfall einschließlich Profiling.« In *Datenschutzrecht. DSGVO und BDSG*, edited by Spiros Simitis, Gerrit Hornung and Indra Spiecker genannt Döhmann. Baden-Baden: Nomos.

Sloane, Mona, Emanuel Moss and Rumman Chowdhury. 2022. »A Silicon Valley love triangle: Hiring algorithms, pseudo-science, and the quest for auditability.« *Patterns* 3 (2): 100425.

Smuha, Nathalie A. 2021. »Beyond the individual: governing AI's societal harm.« *Internet Policy Review* 10 (3).

Teo, Sue A. 2023. »Human dignity and AI: mapping the contours and utility of human dignity in addressing challenges presented by AI.« *Law, Innovation and Technology* 15 (1): 1–39.

Thomsen, Frej K. 2017. »Discrimination.« In *Oxford Research Encyclopedia of Politics* (online), edited by William R. Thompson. New York: Oxford University Press.

Ulgen, Ozlem. 2017. »Kantian Ethics in the Age of Artificial Intelligence and Robotics.« *Questions of International Law (QIL) Zoom-in* 43: 59–83.

Ulgen, Ozlem. 2022. »AI and the Crisis of the Self: Protecting Human Dignity as Status and Respectful Treatment.« In *The Frontlines of Artificial Intelligence Ethics: Human-Centric Perspectives on Technology's Advance*, edited by Andrew J. Hampton and Jeanine A. DeFalco, 9–33. Abingdon, New York: Routledge.

Valcke, Peggy, Damian Clifford and Vilté K. Dessers. 2021. »Constitutional Challenges in the Emotional AI Era.« In *Constitutional Challenges in the Algorithmic Society*, edited by Hans-W. Micklitz, Oreste Pollicino, Amnon Reichman, Andrea Simoncini, Giovanni Sartor and Giovanni De Gregorio, 57–77. Cambridge: Cambridge University Press.

von Ungern-Sternberg, Antje. 2022. »Diskriminierungsschutz bei algorithmenbasierten Entscheidungen.« In *Handbuch Antidiskriminierungsrecht. Strukturen, Rechtsfiguren und Konzepte*, edited by Anna K. Mangold and Mehrdad Payandeh, 1131–80. Tübingen: Mohr Siebeck.

Yeung, Karen. 2019. *Responsibility and AI. A study of the implications of advanced digital technologies (including AI systems) for the concept of responsibility within a human rights framework*. Council of Europe study DGI(2019)05. Council of Europe, Expert Committee on human rights dimensions of automated data processing and different forms of artificial intelligence (MSI-AUT). Strasbourg.

Zarsky, Tal. 2013. »Transparent predictions.« *University of Illinois Law Review* 2013 (4): 1503–69.

Normung und Standardisierung von KI-Systemen aus soziotechnischer Perspektive

Cecilia Colloseus

Der Normungsbedarf von KI-Systemen

Wenn sozialarbeiterische Arbeitsprozesse digital unterstützt und in ihrem Kontext ethisch wie rechtlich begründete Entscheidungen getroffen werden sollen, müssen Systeme, die auf künstlicher Intelligenz (KI) basieren, strengen Standards entsprechen. Gerade wenn es um Einschätzungen von Kindeswohlgefährdung und die Bewertung von Risikopotential geht, ist eine Ausrichtung an fundierten Normen unumgänglich.

Dieser Artikel befasst sich mit der Normung von KI. Normung ist ein wichtiger Schritt, um eine verantwortungsvolle und ethische Anwendung von KI-Technologien zu gewährleisten. Allerdings kann die digitale Operationalisierung in der Normung auch ethische Herausforderungen mit sich bringen, wie z.B. eine ungleiche Beteiligung von Interessengruppen oder die Verstärkung von bestehenden Vorurteilen und Diskriminierungen.

Die genannten Herausforderungen werden in diesem Artikel vor dem Hintergrund des EU-AI-Acts (AI-Act) und der DIN-KI-Normungsroadmap (DIN-KI) betrachtet. Es wird der Frage nachgegangen, ob die Vorgaben der Standardisierung und Normung eine Basis für die Umsetzung der in KAIMo verhandelten Fragestellung, ob ein Algorithmus moralisch kalkulieren kann, bieten können.

KI-Normung auf europäischer und nationaler Ebene – EU-AI-Act und DIN-KI-Normungsroadmap

Der EU-AI-Act ist ein 2021 von der Europäischen Kommission vorgestellter Vorschlag für eine neue Gesetzgebung zur Regulierung von KI. Der Gesetzesentwurf unterteilt KI-Systeme in drei Risikokategorien: Anwendungen, die ein nichtakzeptables Risiko darstellen (und folglich nicht in Umlauf gebracht werden dürfen), Hochrisikoanwendungen mit Regulierungsbedarf und Anwendungen mit geringem Risiko, die keiner weiteren Regulierung bedürfen. Der AI-Act sieht unter anderem die Einführung von verpflichtenden Anforderungen für Hochrisiko-KI-Anwendungen vor sowie die Verpflichtung für Anbietende von KI-Systemen, eine Risikoanalyse durchzuführen und eine umfassende Dokumentation über ihr KI-System bereitzustellen.

Die DIN-KI-Normungsroadmap hingegen ist eine Initiative des Deutschen Instituts für Normung (DIN) zur Entwicklung von Normen und Standards für Künstliche Intelligenz. Die Roadmap wurde 2020 erstmals veröffentlicht und definiert die wichtigsten Handlungsfelder für die Normung von KI-Technologien in Deutschland. Hierzu gehören unter anderem Themen wie Ethik und Sicherheit von KI-Systemen, Datenqualität und -zugang sowie Bildung und Qualifizierung im Bereich Künstlicher Intelligenz.

Beide Initiativen, der EU-AI-Act und die DIN-KI-Normungsroadmap, sind wichtige Schritte zur Entwicklung einer verantwortungsvollen und moralischen Anwendung von KI-Technologien und zeigen das wachsende Bewusstsein für die Bedeutung von KI-Normen und -Standards auf nationaler und internationaler Ebene. Im Kontext von Algorithmen, die im Konfliktfall moralisch kalkulieren sollen, muss ein Framework für eine ethische digitale Operationalisierung in der KI-Normung herausgearbeitet werden, das sowohl theoretisch fundiert als auch praktisch anwendbar ist. Dieses Framework soll dazu beitragen, dass die Normung von KI-Technologien nicht nur technisch korrekt, sondern auch ethisch verantwortungsvoll und gerecht erfolgt.

Es gibt bereits einzelne internationale Standards zu KI: ISO/IEC 22989:2022 (Informationstechnik – Künstliche Intelligenz), ISO/IEC DIS 23894:2022 (Informationstechnik – Künstliche Intelligenz – Risikomanagement), ISO/IEC 23053:2022 (Framework für Systeme der Künstlichen Intelligenz (KI) basierend auf maschinellem Lernen (ML), ISO/IEC DIS 42001 (Information Technology – Artificial intelligence – Management system), ISO/IEC 5259 (Artificial intelligence – Data quality for analytics and machine learning). Für die Implementierung von KI-Systemen in unterschiedlichen betrieblichen

Kontexten muss jedoch auch auf andere Normen und Standards zurückgegriffen werden, wie etwa DIN-Normen zu Ergonomie (z.B. DIN EN ISO 9241–112:2017, DIN EN ISO 9241–110:2020).

Im Folgenden werden Inhalte der DIN-KI-Normungsroadmap und des EU-AI-Acts referiert, die an diesen Anspruch anknüpfen und ein entsprechendes Framework bereitstellen können. Der Fokus liegt hierbei auf Kapitel 4.4 der DIN-KI-Normungsroadmap »Soziotechnische Systeme«. Es sei vorangestellt, dass es in der Normungsroadmap vor allem um die Implementierung von KI-Systemen in Unternehmen geht. Auf Institutionen und Behörden wird nicht gesondert eingegangen und es ist zu prüfen, ob sich die entsprechenden Normen auch auf das im Rahmen von KAIMo untersuchte Feld anwenden lassen.

Soziotechnische KI-Normung

> »Vom Produktentstehungsprozess über Inbetriebnahme und alltäglicher operativer Anwendung bis hin zur Außerbetriebsetzung sind nicht nur der Stand der technologischen Entwicklung sowie der spezifische Anwendungsfall zu berücksichtigen, sondern auch die Grundsätze und Prinzipien einer menschengerechten und partizipativen soziotechnischen Gestaltung. Dieses Erfordernis spiegelt sich bislang meist nicht in den korrespondierenden Normen wider.« (DIN-KI, 162)

Da es im hier verhandelten Kontext um den Gebrauch von Algorithmen in sozialen Konfliktfällen geht, werden die technischen Aspekte der KI-Normung ausgespart und dafür die soziotechnischen Zusammenhänge prominent hervorgehoben. Als soziotechnische Systeme werden solche Systeme bezeichnet, in denen Mensch und Technik miteinander verknüpft sind und in Wechselwirkung zueinander stehen (Zweig et al. 2021, Schlick et al. 2010, Suchman 2009). Auch KI-Technologien sind in diesem Kontext zu betrachten. Der einzelne Mensch, sein organisatorisches Umfeld und die Gesellschaft als Ganzes müssen in den Blick genommen werden, wenn Regeln und Normen hierfür aufgestellt werden sollen.

Im EU AI-Act wird gefordert, dass Hochrisikosysteme mit Funktionen ausgestattet werden, die den Menschen aktiv einbinden. So soll es für alle Betroffenen und Beteiligten ein hohes Maß an Transparenz geben, die menschliche Aufsicht soll gewährleistet und eine Art »Stopptaste« eingerich-

tet werden, die vom Menschen ausgelöst werden kann (EU AI-Act, Artikel 14, 4d). Der Mensch steht also immer im Mittelpunkt. Alle technischen Komponenten müssen sich an den soziotechnischen Anforderungen ausrichten und an der gesamten Entwicklung müssen alle relevanten Akteur*innen beteiligt werden (partizipative Forschung und Design).

In ihrer zweiten Ausgabe adressiert die DIN-KI-Normungsroadmap soziotechnische Aspekte wie die Mensch-Technik-Interaktion, die humane Arbeitsgestaltung sowie Anforderungen an Unternehmensstrukturen und -prozesse in einem eigenen Kapitel (DIN-KI, 153–175). Bereits in den allgemeinen Handlungsempfehlungen zu Beginn wird die Empfehlung ausgesprochen, »[d]en Menschen als Teil des Systems [zu] begreifen, und zwar in allen Phasen des KI-Lebenszyklus.« (DIN-KI, 35). Die Normungsroadmap fordert außerdem, »[g]esellschaftliche und ethische Fragestellungen mithilfe etablierter Modelle […] zu operationalisieren, messbar bereits bei der Entwicklung der Technologie zu verankern und dabei auf dem Stand der Forschung zu Diskriminierungssensibilität aufzubauen« (DIN-KI, 36). Des Weiteren ist vorgesehen, dass eine »adäquate Organisationskultur« etabliert wird. Konkret bedeutet das, dass in den individuellen Unternehmen oder Institutionen die relevanten Akteur*innen sensibilisiert, qualifiziert und in einem geeigneten Change Management im Prozess mitgenommen werden (DIN-KI, 36).

Ein weiterer Anspruch, den die Roadmap in diesem Kontext formuliert, ist, dass die beteiligten Menschen über den gesamten Lebenszyklus von KI-Systemen hinweg mit Prozessen, Methoden und Tools unterstützt werden sollen. Das setzt voraus, dass die vom System betroffenen und daran beteiligten Menschen auch in alle Phasen des KI-Lebenszyklus eingebunden werden. Normungsgremien müssen dafür konkrete Normen erarbeiten, die gewährleisten, dass die notwendigen Transparenzanforderungen und vor allem die menschliche Aufsicht über KI-Systeme eingehalten werden. Auch in der Erarbeitung dieser Normen müssen Menschen aus den unterschiedlichen betroffenen Zielgruppen berücksichtigt werden (z.B. Zivilgesellschaft).

Die soziotechnische Perspektive zeigt auf, dass es bei KI nicht allein um die technische Machbarkeit geht, sondern dass der Kontext der jeweiligen Anwendung beachtet werden muss. Sie ist das »multiperspektivische ›Gegengewicht‹ zu einer rein technikzentrierten Sicht auf KI« (DIN-KI, 156).

Entwicklung von KI-Systemen nach soziotechnischen Kriterien

Laut der KI-Normungsroadmap reicht es nicht aus, die soziotechnische Perspektive nur in der Entwicklung von KI-Systemen einzunehmen. Sie muss vielmehr über den gesamten Lebenszyklus des Systems betrachtet werden (vgl. ISO/IEC 22989:2022, ISO/IEC 23053:2022), da sich KI-Systeme in einem gewissen Rahmen weiterentwickeln. Dieser Rahmen muss in der Designphase abgesteckt werden. Ist das System bereits in Betrieb, kann immer nur ein Ausschnitt des Systemzustands zu einem bestimmten Zeitpunkt abgebildet werden.

In der ersten Phase des KI-Systems, der Initialisierung, werden Ziel und Zweck der Anwendung definiert (vgl. ISO/IEC 22989:2022). Es wird geklärt, welche Anforderungen die KI innerhalb des soziotechnischen Systems erfüllen muss, also welches Problem sie lösen soll, welche Bedürfnisse ihre Zielgruppe hat und welche Erfolgsparameter es gibt. Dabei geht es nicht (nur) um die technische Machbarkeit, sondern »darum, auf Basis einer eingehenden Problemanalyse in einer gegebenen Situation von der Idee zur Entscheidung für ein KI-System zu gelangen und den Entwicklungsprozess anzustoßen« (DIN-KI, 159). Hier müssen auch ethische Aspekte einbezogen werden, z.B. anhand von ethics-by-design-Katalogen wie dem der Bertelsmann Stiftung (Puntschuh und Fetic 2020). Auch das WKIO-Modell der AI Ethics Impact Group liefert eine Methode, vorab definierte ethische Werte zu operationalisieren (VDE, Bertelsmann Stiftung 2020). Teil dieser Initialisierungsphase ist außerdem eine erste Risikoanalyse, die die soziotechnischen Folgen aus Sicht mehrerer Stakeholder identifiziert (ISO/IEC 23894:2022). Hier geht es nicht nur um technische und rechtliche Fragen, sondern auch und vor allem um ethische und soziale Folgen. So müssen die Fragen geklärt werden, welche Grundrechte oder -werte durch den Einsatz der Software potenziell berührt werden, was die beabsichtigten Auswirkungen der Software sind, wer vom Einsatz des algorithmischen Assistenzsystems betroffen ist, welche potenziellen Auswirkungen der Einsatz der Software auf die unterschiedlichen Stakeholder, Gesellschaft, Wirtschaft oder Umwelt hat, welche Risiken durch mögliche Fehler bei der Entwicklung oder dem Einsatz der Software entstehen können und welche Szenarien hier denkbar sind (Puntschuh und Fetic 2020).

Participation is Key

Bevor die Entscheidung für ein KI-System getroffen wird, muss definiert werden, welche Personen (-gruppen) überhaupt davon betroffen sein werden und welche Bedürfnisse sie haben. Diese Personen müssen an der Entwicklung und Implementierung des Systems unbedingt beteiligt werden (DIN-KI, 160). Nur, wer die Zielgruppe kennt, kann ein System entwickeln, das divers ist und keine stereotypischen Vorstellungen von Menschen (-gruppen) reproduziert. Ein möglicher Ansatz für die Einbindung von unterschiedlichen Betroffenen ist der Participatory-Design-Ansatz (vgl. Simonsen und Robertson 2013). Gemeinsam werden Szenarien erarbeitet, die sich im jeweiligen Kontext mit Implementierung des KI-Systems ergeben können. So werden z.B. gemeinsam Prototypen entwickelt oder bestimmte Situationen simuliert. Nutzende erhalten auf diese Art ein besonderes Mitspracherecht, ohne selbst entsprechende technische Fachkenntnisse haben zu müssen. Im folgenden Entwicklungsprozess werden sie an der Evaluation des Systems beteiligt. Hier bieten sich Ansätze von XAI (Explainable AI) an, die über Visualisierung einen Zugang zu den zugrunde liegenden Softwarelogiken ermöglichen oder kritische Entscheidungspunkte offenlegen. Grundsätzlich wäre die idealtypische Forderung bei der Planung und Gestaltung einer KI-Lösung, eine Vertretung aller Stakeholder zu beteiligen.

Die ISO/IEC 22989:2022 unterteilt Stakeholder in: »AI-Provider«, »AI-Producer«, »AI-Customer«, »AI-Partner*innen«, »AI-Subject« und »Relevant Authorities«. Einzubeziehen sind z.B. Expert*innen mit Domänenwissen (KI-Expert*innen, Data Scientists, Informatiker*innen usw., Prozessgestaltende, Usability- Expert*innen, Produktgestaltende usw., Softwaretester*innen, Ergonom*innen, Psycholog*innen usw., Sicherheitsexpert*innen in der jeweiligen Domäne, Expert*innen für Ethik, Diversity, Fairness usw.), Expert*innen aus den betroffenen Fachabteilungen, Nutzende des KI-Systems, Interessensvertretungen von Betreibenden und Nutzenden, Entscheider*innen über den Einsatz der KI-Lösung, Vertreter*innen der Zivilgesellschaft, aber auch »überraschende« Stakeholder, also solche Personen, die nicht unmittelbar mit dem jeweiligen System zu tun haben, aber mittelbar betroffen sein können. Wie und in welchem Kontext die jeweiligen Stakeholder eingebunden werden, ist abhängig vom Zeitpunkt im Projektlebenszyklus, also während der Zielfindung, während der Planung und Gestaltung, während der Inbetriebnahme, im laufenden Betrieb bzw. im kontinuierlichen Verbesserungsprozess.

Die Kommunikation muss dabei immer zielgruppengerecht und inklusiv sein. So ist die Kommunikation zwischen den Expert*innen mit Domänenwissen untereinander, zwischen den Expert*innen mit Domänenwissen und den Nutzenden, zwischen Nutzenden und Technik, zwischen Expert*innen und sonstigen Beteiligten jeweils unterschiedlich. Um die Beteiligung gut durchzuführen, können einschlägige Normen und Standards (z.B. VDI-MT 7001:2021) herangezogen werden. Zu beachten ist darüber hinaus, dass die Beschäftigten entsprechende Kompetenzen erwerben müssen, um mit dem System richtig umgehen zu können. Entsprechende Schulungen müssen frühzeitig durchgeführt und am Wissensstand der Beschäftigten ausgerichtet werden.

KI-Systeme – Werkzeuge oder Agenten?

Wenn IT-Systeme (insbesondere KI) soziotechnisch gestaltet werden, müssen sie geeignet sein, (Arbeits-)Aufgaben von Menschen in unterschiedlichen Rollen und im Nutzungskontext zu unterstützen. Das bedeutet zum Beispiel, dass die Schnittstellen ergonomisch gestaltet werden müssen. Die Nutzenden sollen vom KI-System dabei unterstützt werden, Aufgaben effektiv, effizient und zufriedenstellend zu erledigen. Dabei müssen zunächst die Bedürfnisse der Nutzenden erkannt und analysiert werden. Methoden der partizipativen Sozialforschung sind hier das Mittel der Wahl.

In der Normungsroadmap wird konstatiert, dass KI-Systeme nicht nur als technische Werkzeuge wahrgenommen werden. Vielmehr sollen sie als »eine neue Klasse von Agenten in der Organisation« (Raisch und Krakowski 2020) betrachtet werden. Die Interaktion zwischen Menschen und diesen neuen nichtmenschlichen Agenten wird in Autonomiestufen unterteilt. So soll die KI etwa nur dann eine Aufgabe erledigen können, wenn ein Mensch die Bestätigung dafür gibt, oder es muss möglich sein, dass der kontrollierende Mensch ein Veto gegen die Entscheidungen einer KI einlegt. Auf einer höheren Autonomiestufe handelt die KI vollständig autonom und der Mensch wird nur dann informiert, wenn er konkret nachfragt. Die höchste Stufe von Autonomie ist dann erreicht, wenn die KI handeln kann, ohne den Menschen einzubeziehen.

Bislang wurden in Konzepten wie Ergonomics/Human Factors (EHF) statische technische Systeme betrachtet (z.B. stationäre Maschinen). KI-Systeme sind jedoch inhaltlich und zeitlich dynamisch. Deshalb muss das EHF-Gestal-

tungskonzept erweitert werden, damit die Dynamik von Schnittstellen, Funktionsweisen und Auswirkungen auch für Menschen passend gestaltet werden können. So werden etwa auch menschliche Eigenschaften wie Empathie mit einbezogen (André und Bauer 2021, Höddinghaus et al. 2021, Brynjolfsson et al. 2018, Moray 1989). Werden Systeme nach soziotechnischen Grundsätzen gestaltet, werden Technologieeinsatz und Organisation gemeinsam optimiert (»joint optimization«) (vgl. Cherns 1976&1987, Ulich 2013).

Die KI als gesellschaftliche Akteurin

Neben der Organisation und dem (individuellen) Menschen, spielt in soziotechnischen Systemen auch die umgebende Gesellschaft eine entscheidende Rolle. Sie bildet die Schnittstelle von Individuum und Organisation. Bestehende Ungleichheitsverhältnisse oder Diskriminierung, die innerhalb der Gesellschaft bestehen, können sich in KI-Anwendungen verfestigen. Technologien, die menschliche Intelligenz nachahmen sollen, können deshalb niemals objektiv oder neutral sein (Benjamin 2019), da sie eben immer in eine von bestimmten Werten – und damit einhergehend: sozialen Herausforderungen – geprägten Gesellschaft eingebettet sind. Nutzen Menschen solche Technologien, beeinflussen sie diese und umgekehrt (Suchman 2007). Besonders deutlich wird das beim Maschinellen Lernen: Hier reagieren Softwareprogramme dynamisch und adaptiv auf ihre Nutzer*innen (Pentenrieder et al. 2020). Betrachtet man Mensch und Maschine auf diese Weise, wird die bisherige Konzeption von autonomen und strikt trennbaren Entitäten in Frage gestellt. Menschliche und maschinelle Akteur*innen ergeben durch Kollaboration und Interaktion ein Ganzes (Suchman 2007).

Training des KI-Systems: Das Datenset

Wird ein konsistentes soziotechnisches Mensch-Technik-Organisation-Modell verwendet, müssen zentrale Fragen der Einführung, Nutzung und Folgeabschätzung von KI bearbeitet werden (Huchler et al. 2020, Stowasser und Suchy 2020). Die erste Frage betrifft dabei die Daten, auf deren Grundlage die KI entwickelt wird und den Zweck, dem sie dienen soll. Als zweites stellt sich die Frage, wie das menschliche Verhalten durch den Einsatz der KI beeinflusst wird z.B. Autonomie oder Entscheidungsdilemmata. Die dritte Frage betrifft

das Verhältnis zwischen KI und menschlichen Bedürfnissen und die vierte Frage die systemischen Folgewirkungen. Hier geht es sowohl um die Wirkungen innerhalb des Systems als auch in den Subsystemen, der Systemumwelt und Gesellschaft (DIN-KI, 155). Zur ersten Frage sei erläutert: KI-Systeme müssen anhand spezifischer Daten trainiert werden, um in den betrieblichen Einsatz überführt werden zu können. Die Auswahl der Trainings-, Validierungs- und Testdaten muss dabei so erfolgen, dass Diskriminierung vermieden wird. Außerdem müssen die Daten auf ihre Qualität hinsichtlich des geplanten Einsatzes überprüft werden: Sind genug Daten vorhanden, sind sie konsistent, sind die Daten aktuell, befinden sich falsche Daten im Set etc.? Auswahl, Training, Verifizierung und Validierung der verwendeten Datensätze und die Testung der KI-Lösung sind adäquat zu dokumentieren.

Es ist hinlänglich bekannt, dass Menschen systematisch Entscheidungsfehler machen (Kahnemann und Tversky 1982). In dieser Hinsicht sind ihnen KI-Systeme ähnlich, denn auch sie können in ihrer Entwicklung und im Einsatz Bias-Effekte oder Entscheidungsfehler bezüglich der Fairness verursachen. Als »Bias« werden unerwünschte Verzerrungen bezeichnet, die entweder bereits bei der Erhebung der Datensätze oder durch die Selektion bzw. Art ihrer Verarbeitung aufkommen können. Diese Bias-Effekte sind jedoch bedingt durch menschliche Designentscheidungen (z.B. Datenbank und Logik) und die dem jeweiligen Einsatzgebiet zugrunde liegenden Vorannahmen. Die Rechenschaftspflicht und die Gewährleistung von Fairness liegen also beim Menschen. Bei Bias-Effekten sind die Faktoren Unsicherheit und Risiko zu beachten. Als Risiko wird bezeichnet, wenn ein bestimmter bekannter Umweltzustand mit einer empirisch ermittelten Wahrscheinlichkeit eintreffen kann. Risiko ist also quantifizierbar und potentiell steuerbar. Unsicherheit hingegen bedeutet, dass weder die möglichen Umweltzustände noch die mögliche Eintrittswahrscheinlichkeit bekannt sind (DIN-KI, 156). Algorithmen für Risikosituationen und -abschätzung wägen das Risiko in Form von Eintrittswahrscheinlichkeiten und gewünschten Optimierungslevels ab (Kahnemann und Tversky 1982). Dabei wird deutlich, dass die menschliche Wahrnehmung bei der Analyse, Gestaltung und Bewertung von KI-Systemen eine entscheidende Bedeutung hat. Zur Einschätzung der Risiken gehört auch, Ansprüche an Transparency und Rechenschaftspflicht zu formulieren. Welche Informationen müssen für wen offengelegt werden? Mit welcher technischen Tiefe müssen Informationen angereichert werden, um gleichzeitig hilfreich und verständlich zu sein? Wer ist rechenschaftspflichtig im Schadensfall? Im Anschluss an das Risiko-Assessment müssen Kosten, Aufwand

und Ressourcen für die Umsetzbarkeit der Anwendung ermittelt werden (vgl. ISO/IEC 22989:2022).

Soziale Nachhaltigkeit im soziotechnischen Kontext

KI-Anwendungen müssen bestimmten Nachhaltigkeitskriterien genügen und parametrisierbar sein in Bezug auf quantitative Zielsetzungen.

> »Im Hinblick auf die Entwicklung, Nutzung und den Einsatz von KI-Systemen bedeutet Nachhaltigkeit vor allem, dass die Würde des Menschen respektiert wird, keine Menschen ausgeschlossen, benachteiligt oder diskriminiert werden und die menschliche Autonomie und Handlungsfreiheit durch KI-Systeme nicht eingeschränkt werden dürfen. In einer erweiterten Perspektive auf Nachhaltigkeit bedeutet soziale Nachhaltigkeit auch, dass neben körperlicher Unversehrtheit und menschenwürdigen Lebensbedingungen auch die Fähigkeit, auf menschliche Art und Weise zu denken, zu argumentieren und zu handeln, nicht eingeschränkt werden sollte. Hier zeigt sich schon, dass ein umfassendes Verständnis von sozialer Nachhaltigkeit sehr weitreichende Konsequenzen für die Gestaltung von KI-Systemen hat.« (Rohde et al. 2021).

Gesetze und Normen können nicht jedes Detail regeln. Subsidiäre Aushandlungssysteme, z.B. auf betrieblicher Ebene und individuelle Entscheidungsrechte müssen hier ergänzend eingesetzt werden. Muhammad (2022) konstatiert, dass KI-Systeme Individuen und Gruppen Schaden zufügen können und etabliert verschiedene Typen von Fehlern, die in diesem Kontext passieren können:

- »Vergabe-Fehler«: Das System hält Möglichkeiten, Ressourcen oder Informationen zurück oder stellt sie unfair zur Verfügung.
- »Servicegüte-Fehler«: Das System arbeitet nicht für alle Gruppen ähnlich gut.
- »Repräsentations-Fehler«: Die Entwicklung oder die Verwendung eines Systems repräsentiert Gruppen unterschiedlich.
- »Stereotyp-Fehler«: Das System reproduziert und verstärkt Stereotypen, indem beispielsweise stereotypische Charakteristika unreflektiert allen Angehörigen einer Gruppe zugewiesen werden.

- »Verunglimpfungs-Fehler«: Das System wird aktiv abwertend oder beleidigend.
- »Prozess-Fehler«: Das System trifft Entscheidungen aufgrund von Charakteristika, die nicht für die Aufgabe relevant sein sollten (DIN-KI, 156f.)

Im Datenschutz müssen die Grundprinzipien Zweckfeststellung bzw. Zweckbindung von Daten, Erforderlichkeit der Datensammlung, Transparenz, Datenvermeidung und Datensparsamkeit eingehalten werden. Diese Prinzipien gelten auch für die Gestaltung soziotechnischer Systeme.

Für ethische Betrachtungen wurden Gütesiegel vorgeschlagen, die auf Wertanalyseverfahren aus einer Kombination von Zielkriterien, Indikatoren und messbaren Größen beruhen. Die Bewertung dieser normativen Anforderungen sollten sich auf technische Prüfungen stützen können.

Anwendungsspezifische Anforderungen bringen normative Grundsätze in die konkrete Anwendung und fügen spezielle Einsatzanforderungen hinzu. Sie bilden die Grundlage der Risikoeinstufung gemäß AI Act, greifen dabei relevante ethische Aspekte auf, nutzen das New Legislative Framework, um komponentenweise die Konformitätsvermutung von Herstellenden zu unterstützen und formulieren im Grundsatz Anforderungen an das gesamte technische System, in das KI eingebettet ist. Hierbei geht es um die vertikale Bewertung von KI, bei der geprüft wird, ob die KI für einen bestimmten Einsatzzweck geeignet ist (DIN-KI, 128).

Funktionsteilung Mensch-Technik

In der Funktionsteilung zwischen Mensch und KI-System gilt das Primat der (Arbeits)Aufgabe. Die Gestaltung der Aufgabe steht also am Anfang des Gestaltungsprozesses und ordnet ihr die Gestaltung der Ausführungsbedingungen unter (Hacker und Sachse 2014, Ulich 2011, DIN EN 614–2:2008). Der Autonomiegrad des KI-Systems wird repräsentiert durch die Funktionsteilung. Die einschlägigen Normen sind dahingehend zu prüfen, ob sie die verschiedenen Autonomiegrade berücksichtigen. Für die Funktionsteilung wird die von Fitts (1951) entwickelte HABA MABA-Liste (= »humans are better at« – »machines are better at«) herangezogen. Es kann aber keine starre Funktionszuteilung zwischen Mensch und Technik vorgenommen werden, da diese Subsysteme nicht mechanistisch zusammenwirken. Außerdem pauschaliert eine solche verkürzende Dartstellungsweise Fertigkeiten, Fähigkeiten und Wissen

des Subsystems Mensch und beachtet die Dynamik und Weiterentwicklung der Subsysteme nicht. Auch wird die Lebenszyklusperspektive für beide Subsysteme nicht berücksichtigt. Die Funktionsteilung kann sich aber abhängig von der Situation dynamisch ändern, z.B. wenn der Mensch in einer Gefahrensituation eingreifen muss. Zu dieser Adaptivität gibt es aktuell noch keine Normung.

In der dynamischen Funktionsallokation kann es zu »Ironien der Automatisierung« kommen (vgl. Bainbridge 1987). Das bedeutet, dass mit der Automatisierung die Systemkomplexität steigt und neue Aufgaben der Überwachung, Steuerung und Korrektur entstehen, für die der Mensch nicht ausreichend qualifiziert ist. Wenn Funktionsteilung und Automatisierung gestaltet werden, muss diese »Ironie« berücksichtigt werden und in die relevanten Normen und Standards einfließen.

Entscheidend ist außerdem das Menschenbild:

> »Wird der Mensch vom Entwickelnden als Fehlerquelle gesehen, so wird die Gestaltung tendenziell versuchen, den Einfluss des Menschen im KI-System weitgehend zu reduzieren. Wird das KI-System hingegen als Unterstützung für den Menschen betrachtet, so wird die Funktionsteilung eher komplementär erfolgen.« (DIN 166)

Im Interesse einer menschenzentrierten KI-Nutzung sollte dem Leitbild einer komplementären Funktionsteilung der Vorzug gegeben werden.

Um ein KI-System sinnvoll nutzen zu können, müssen auch die Organisation und Prozesse, in die es eingebettet ist, auf eine bestimmte Weise gestaltet werden: So muss ein Vertrauen in die Automation etabliert werden und die Potentiale für Belastung und Beanspruchung durch die Nutzung des KI-Systems (z.B. Technikstress) erhoben werden (vgl. DIN-EN-ISO-10075-Reihe). Es müssen systemische Effekte mitbedacht werden (z.B. Aufschaukelungseffekte durch den Eingriff des Menschen in das KI-System) und die veränderte Risikokompensation der Nutzenden sowie deren Folgen beim (unbemerkten) Ausfall des Systems.

Die Qualifizierung der Nutzenden, die partizipative Gestaltung und ein geeignetes Change Management sind hier entscheidend. Es ist zu prüfen, inwieweit diese Aspekte in den relevanten Normen und Standards (z.B. DIN EN ISO 27500:2017, VDI/VDE-MT 7100, DIN EN ISO 9001:2015) abgebildet sind.

Grundsätzlich sind bezogen auf Mensch-Technik-Interaktionen in soziotechnischen Systemen drei hierarchisch strukturierte Schnittstellen mit

jeweils darauf bezogenen Gestaltungsprinzipien von besonderem Interesse: Aufgabenschnittstelle, z.B. nach DIN EN 614–2:2008, Reihe DIN EN ISO 11064:2011, Interaktionsschnittstelle, z.B. nach DIN EN 894–1:2009, DIN EN ISO 9241–11:2018, DIN EN ISO 9241- 110:2020, ISO/IEC 29138–1:2018, Informationsschnittstelle, z.B. nach VDI/VDE 3850–1:2014, ISO 9241–112:2017).

Implementierung der KI-Lösung im soziotechnischen System

Ist die Entscheidung für den KI-Einsatz gefallen und die Grob- und Feinplanung der KI-Lösung abgeschlossen, muss ihre Inbetriebnahme anhand von Projektmanagement-Normen geplant werden (z.B. DIN ISO 21500:2016, Reihe DIN 69901, Reihe DIN 69909). Dabei muss geprüft werden, ob KI-Projekte hinsichtlich des Projektmanagements Besonderheiten aufweisen, die bei Bedarf in der Normung abzubilden sind.

Die weiteren Entwicklungsschritte erfolgen iterativ. So können Informationen, die im Betrieb des Systems neu hinzukommen, eine Rückkehr zu den Schritten in der Initialisierungsphase erforderlich machen. Der Prozess orientiert sich an den Schritten »Design und Entwicklung« und »Verifikation und Validierung« der ISO/IEC 22989:2022. Die in dieser Phase entwickelte Groblösung des KI-Systems wird für die Inbetriebnahme vorbereitet. Alle Beteiligten, etwa die Betreibenden des KI-Systems, spätere Nutzende des KI-Systems, Interessensvertretungen von Betreibenden und Nutzenden, Vertretende der Zivilgesellschaft, aber auch »überraschende Stakeholder«, müssen in dieser Phase einbezogen werden. Nur so kann die Umsetzung von ergonomischen Grundsätzen und Prinzipien sowie eine gebrauchstaugliche Gestaltung von Produkten und Arbeitsmitteln gewährleistet werden.

Mensch, Technik, Organisation

Soziotechnische Systeme werden immer von sachlichen (technisch-organisatorischen) und menschlichen (persönlichen) Gegebenheiten beeinflusst (z.B. DIN EN ISO 6385:2016). Deshalb sind Mensch, Technik und Organisation stets in ihrer gegenseitigen Abhängigkeit und ihrem Zusammenwirken zu reflektieren (MTO-Konzept). Die Gestaltungsdimensionen eines KI-Systems ergeben sich daher also aus den Elementen Mensch, Technik und Organisation sowie aus deren Schnittstellen (Mensch-Technik, Mensch-Organisation und Organi-

sation-Technik). Aus den identifizierten Gestaltungsdimensionen resultieren verschiedene Normen und Standards, die für die Planung und Gestaltung des jeweiligen Systems heranzuziehen sind.

Bei der Planung und Gestaltung jedes KI-Systems ist zwingend eine Analyse des zugrunde liegenden soziotechnischen Systems durchzuführen. Im Arbeitskontext ist das soziotechnische System das »Arbeitssystem« (vgl. Ulich 2013). Laut DIN EN ISO 6385:2016 ist das Arbeitssystem ein »System, welches das Zusammenwirken eines einzelnen oder mehrerer Arbeitender/Benutzer*innen mit den Arbeitsmitteln umfasst, um die Funktion des Systems, innerhalb des Arbeitsraumes und der Arbeitsumgebung unter den durch die Arbeitsaufgaben vorgegebenen Bedingungen, zu erfüllen«.

Ethische Aspekte sind bei der Planung und Gestaltung des soziotechnischen Systems stets zu beachten und für den gesamten Lebenszyklus des KI-Systems zu gestalten. Dazu zählen z.B. Transparency, Accountability, Privacy, Justice, Reliability und Sustainability (z.B. AI Ethics Label der AI Ethics Impact Group, VDE 2020).

Die einschlägigen Normen bzw. Standards zu KI (z.B. ISO/IEC 22989, ISO/IEC 42001, DIN SPEC 92001 Reihe, DIN SPEC 92001–2:2020, DIN SPEC 92001–3, ISO IEC 25059:2022), Ergonomie & Organisation (z.B. DIN EN ISO 6385:2016, DIN EN ISO 26800:2011, DIN EN ISO 9241 Reihe, Reihe, DIN EN ISO 27500:2017, VDI/VDE-MT 7100) und Ethik (VDE SPEC 90012, IEEE 7000 Serie, ISO IEC/TR 24028) berücksichtigen die resultierenden Anforderungen aus der soziotechnischen Gestaltung eines KI-Systems noch nicht hinreichend und lassen oft die Wechselwirkungen zwischen Mensch, Technik und Organisation außer Acht.

Überprüfung des KI-Systems im Regelbetrieb

In regelmäßigen Abständen muss überprüft und entschieden werden, ob die gewünschte Funktionsweise an geänderte Rahmenbedingungen angepasst werden muss. Im Betrieb können Echtdaten (je nach Anwendungsfeld anonymisiert oder pseudonymisiert) gesammelt und für die relevanten Akteur*innen im soziotechnischen System verständlich und transparent aufbereitet werden. So kann das System kontinuierlich verbessert werden.

Auch in diese kontinuierliche Evaluierung und Anpassung des Systems sollten die Betroffenen eingebunden werden. Ihre Erfahrungen sind die Grundlage für nötige Verbesserungen (Stowasser und Suchy 2020). Da-

für wird eine technische Lösung mit einem Transparency-by-Design- bzw. Transparency-by-Default-Ansatzes benötigt, die Menschen ermöglichen, den Überblick zu behalten. Das kann als Modul im System erfolgen oder durch ein eigenständiges Command Tool.

Die Stopptaste

Personen, die innerhalb des Systems eine Expert*innenrolle einnehmen, werden unter dem Begriff High Level Expert Group (HLEG) zusammengefasst. Sie gewährleisten die menschliche Aufsicht über das KI-System und werden dann als »Human-in-the-Loop (HITL, im Entscheidungszyklus der KI involviert), Human-on-the-Loop (HOTL, beim Design der KI und im Monitoring involviert) oder »Human in Command« (HIC, soll die Gesamtaktivität inklusive breitere ökonomische, soziale, rechtliche und ethische Auswirkungen überblicken können)« (DIN-KI, 169) bezeichnet. Der HIC wird im EU-AI-Act prominent vorgestellt und soll v.a. für Hochrisikosysteme, über geeignete Interventionsmöglichkeiten verfügen. Dazu zählt auch das Auslösen einer »Stopptaste« für die KI (Heesen et al. 2021). »Stopptaste« bedeutet nicht, eine KI-Prozedur bei Zweifeln zu unterbrechen, sondern »die Möglichkeit, einer durch KI getroffenen Entscheidung nicht zu folgen oder die KI-Nutzung für einen bestimmten Zeitraum auszusetzen und stattdessen Menschen entscheiden zu lassen«. (DIN-KI, 169)

Menschliche Interventionen sollten immer möglich sein. Etwa sollten Menschen Ausnahmen von den Entscheidungen der KI treffen können, oder Parameter des Systems (Schwellenwerte, Eingangsgrößen) rekonfigurieren. Der Ansatz »keep the human in the loop« betrachtet einzelne Individuen im Verhältnis zur KI. Demgegenüber steht das Gestaltungsprinzip »keep the organization in the loop«. Es sollen also auch die Interaktion der relevanten Stakeholder betrachtet und laufend optimiert werden (Hermann 2020).

Erklärbarkeit

Es ist von großer Bedeutung, dass KI-Systeme erklärt werden können (Explainable AI, XAI). Nutzende sollten verstehen können, welche Inputs zu welchen Outputs führen, welche Aspekte im Vorhinein festgelegt werden können und welche durch Erklärbarkeitsmetriken identifiziert oder im Nachhinein

über Zielkorridore ermittelt werden. Bei der Entwicklung von soziotechnischen Systemen werden Ziele und Maßnahmen unter Berücksichtigung von Folgeabschätzungen definiert, aber nicht alle Entscheidungen können im Voraus getroffen werden, da sich Rahmenbedingungen ändern können und unbeabsichtigte Effekte auftreten können. Veränderungen des Datensatzes im Betrieb des KI-Systems können durch Methoden im Bereich »Drift« erkannt werden und sollten eine Standardfunktion im KI-System sein. Eine kontinuierliche Überprüfung und Bewertung der Ziele und Entscheidungen in Bezug auf das KI-System ist notwendig und sogar verpflichtend für Hochrisikosysteme gemäß AI Act, Art. 61. Dabei müssen wichtige Fragen beantwortet werden, wie z.B. ob weitere Ziele berücksichtigt werden müssen, ob die Funktionsweise noch korrekt ist und ob Probleme erkannt werden können. »Near Misses« (beinahe Ausfälle) sind oft wertvolle Einblicke in das System, wenn es an seinen Grenzen betrieben wurde. Es ist wichtig, Berichtsstrukturen über Ausfälle oder fast-Ausfälle zu schaffen, um KI-Systeme zu verbessern und zukünftige Systemresilienz sicherzustellen oder Risiken zu simulieren. Es ist auch ratsam, regelmäßig zu überprüfen, ob es grundlegende Änderungen in der KI-Technologie gibt, die die eigene Lösung verbessern oder ersetzen könnten.

Zusammenfassung und Ausblick

Wie der EU-AI-Act und die DIN-Normungsroadmap KI zeigen, ist die Notwendigkeit von Normung und Standardisierung im Kontext von KI von den zuständigen Gremien erkannt worden. Wie bei allen technischen Entwicklungen zuvor, wird es auch für KI-Anwendungen unvermeidbar sein, Zertifizierungen und entsprechende Audits durchzuführen. KI-Systeme, die dabei helfen sollen, im Konfliktfall moralisch zu kalkulieren, fallen in die Kategorie »Hochrisikosystem« und müssen strengen Normen und Standards entsprechen. Diese müssen mit Bedacht festgesetzt werden. Wie in diesem Beitrag dargelegt wurde, ist es entscheidend, vor Einführung eines KI-Systems die tatsächlichen Bedarfe zu identifizieren. Allen voran muss die Frage geklärt werden, ob der Einsatz einer KI wirklich notwendig und sinnvoll ist, oder ob eine andere Lösung gefunden werden kann. Fällt die Entscheidung zugunsten der KI aus, müssen die davon Betroffenen in jeden Schritt der Entwicklung der individuellen KI-Lösung einbezogen werden.

Literatur

André, Elisabeth, and Wilhelm Bauer. 2021. *Kompetenzentwicklung für Künstliche Intelligenz – Veränderungen, Bedarfe und Handlungsoptionen*. Whitepaper aus der Plattform Lernende Systeme, München. DOI: https://doi.org/10.48669/pls_2021-2.

Bainbridge, Lisanne. 1987. »Ironies of Automation.« In: *New Technology and Human Error,* edited by Rasmussen, Jens, Keith Duncan, and Jacques Leplat. John Wiley, New York.

Benjamin, Ruha. 2019. »Race after Technology: Abolitionist Tools for the New Jim Code.« *Social Forces* 98 (4). DOI: https://doi.org/10.1093/sf/soz162.

Brynjolfsson, Erik, Tom Mitchell, and Daniel Rock. 2018. »What Can Machines Learn, and What Does It Mean for Occupations and the Economy?« *AEA Papers and Proceedings* 108: 43–47. DOI: https://doi.org/10.1257/pandp.20181019.

Cherns, Albert. 1976. »The Principles of Sociotechnical Design.« *Human Relations* 29 (8): 783–92. DOI: https://doi.org/10.1177/001872677602900806.

Cherns, Albert. 1987. »Principles of Sociotechnical Design Revisted.« *Human Relations* 40 (3): 153–61. DOI: https://doi.org/10.1177/001872678704000303.

Deutsches Institut für Normung: Deutsche Normungsroadmap Künstliche Intelligenz. Ausgabe 2. 2023. (DIN-KI).

DIN EN 614–2:2008, Sicherheit von Maschinen – Ergonomische Gestaltungsgrundsätze – Teil 2: Wechselwirkungen zwischen der Gestaltung von Maschinen und den Arbeitsaufgaben; Deutsche Fassung EN 614–2:2000+A1:2008.

DIN EN 894–1:2009, Sicherheit von Maschinen – Ergonomische Anforderungen an die Gestaltung von Anzeigen und Stellteilen – Teil 1: Allgemeine Leitsätze für Benutzer-Interaktion mit Anzeigen und Stellteilen; Deutsche Fassung EN 894–1:1997+A1:2008.

DIN EN ISO 26800:2011, Ergonomie – Genereller Ansatz, Prinzipien und Konzepte (ISO 26800:2011); Deutsche Fassung EN ISO 26800:2011.

DIN EN ISO 9001:2015, Qualitätsmanagementsysteme – Anforderungen (ISO 9001:2015); Deutsche und Englische Fassung EN ISO 9001:2015.

DIN 69909 (alle Teile), Multiprojektmanagement – Management von Projektportfolios, Programmen und Projekten.

DIN EN ISO 6385:2016, Grundsätze der Ergonomie für die Gestaltung von Arbeitssystemen (ISO 6385:2016); Deutsche Fassung EN ISO 6385:2016.

DIN EN ISO 9241 (alle Teile), Ergonomie der Mensch-System-Interaktion.

DIN ISO 21500:2016, Leitlinien Projektmanagement (ISO 21500:2012).

DIN 69901 (alle Teile), Projektmanagement – Projektmanagementsysteme.

DIN EN ISO 9241–112:2017, Ergonomie der Mensch-System-Interaktion – Teil 112: Grundsätze der Informationsdarstellung (ISO 9241–112:2017); Deutsche Fassung EN ISO 9241–112:201.

DIN EN ISO 27500:2017, Die menschzentrierte Organisation – Zweck und allgemeine Grundsätze (ISO 27500:2016); Deutsche Fassung EN ISO 27500:2017.

DIN EN ISO 9241–11:2018, Ergonomie der Mensch-System-Interaktion – Teil 11: Gebrauchstauglichkeit: Begriffe und Konzepte (ISO 9241–11:2018); Deutsche Fassung EN ISO 9241–11:2018.

DIN SPEC 92001–1:2019, Künstliche Intelligenz – Life Cycle Prozesse und Qualitätsanforderungen – Teil 1: Qualitäts-Meta-Modell. Letzter Zugriff: 1. April 2023. https://www.beuth.de/de/technische-regel/din-spec-92001-1/303650673.

DIN SPEC 92001–2:2020, Artificial Intelligence – Life Cycle Processes and Quality Requirements – Part 2: Robustness.

DIN SPEC 92001–3, Künstliche Intelligenz – Life Cycle Prozesse und Qualitätsanforderungen – Teil 3: Explainability.

DIN EN ISO 9241–110:2020, Ergonomie der Mensch-System-Interaktion – Teil 110: Interaktionsprinzipien (ISO 9241–110:2020); Deutsche Fassung EN ISO 9241–110:2020.

DIN EN ISO 10075 (alle Teile), Ergonomische Grundlagen bezüglich psychischer Arbeitsbelastung.

Europäische Kommission. 2021. *Vorschlag für eine Verordnung des Europäischen Parlaments und des Rates zur Festlegung harmonisierter Vorschriften für künstliche Intelligenz (Gesetz über künstliche Intelligenz) und zur Änderung bestimmter Rechtsakte der Union*. (EU-AI-Act). Accessed April 1, 2023. https://eur-lex.europa.eu/legal-content/DE/TXT/?uri=CELEX :52021PC0206.

Fitts, Paul M. 1951. *Human engineering for an effective air-navigation and traffic-control system*. Washington, DC: National Research Council.

Hacker, Winfried, and Pierre Sachse. 2013. *Allgemeine Arbeitspsychologie: Psychische Regulation von Tätigkeiten*. Göttingen.

Jessica Heesen, Jörn Müller-Quade, Stefan Wrobel et al. 2021. *Kritikalität von KI-Systemen in ihren jeweiligen Anwendungskontexten – Ein notwendiger, aber nicht hinreichender Baustein für Vertrauenswürdigkeit*. Whitepaper aus der Plattform Lernende Systeme, München.

Herrmann, Thomas. 2020. *Socio-technical design of hybrid Intelligence systems- the case of predictive maintenance*. Accessed April 1, 2023. https://link.springer.com/chapter/10.1007/978-3-030-50334-5_20.

Höddinghaus, Miriam, Dominik Sondern, and Guido Hertel. 2021. »The Automation of Leadership Functions: Would People Trust Decision Algorithms?« *Computers in Human Behavior* 116 (March): 106635. DOI: https://doi.org/10.1016/j.chb.2020.106635.

Huchler, Norbert et al. 2020. *Kriterien für die Mensch-Maschine-Interaktion bei KI. Ansätze für die menschengerechte Gestaltung in der Arbeitswelt*. Accessed April 1, 2023. https://www.plattform-lernende-systeme.de/files/Downloads/Publikationen/AG2_Whitepaper2_220620.pdf.

IEEE 7000:2021, IEEE Standard Model Process for Addressing Ethical Concerns during System Design. Accessed April 1, 2023. https://standards.ieee.org/ieee/7000/6781/#.

IEEE 7001:2021, Standard for Transparency of Autonomous Systems.

IEEE 7002:2022, Standard for Data Privacy Process.

IEEE 7007:2021, Ontological Standard for Ethically driven Robotics and Automation Systems.

IEEE 7005:2021, Transparent Employer Data Governance.

ISO/IEC 29138–1:2018, Informationstechnik – Barrierefreie Benutzungsschnittstellen – Teil 1: Barrierefreiheitserfordernisse der Benutzer.

ISO/IEC 22989:2022, Informationstechnik – Künstliche Intelligenz – Konzepte und Terminologie der Künstlichen Intelligenz.

ISO/IEC DIS 23894:2022, Informationstechnik – Künstliche Intelligenz – Risikomanagement.

ISO/IEC 23053:2022, Framework für Systeme der Künstlichen Intelligenz (KI) basierend auf maschinellem Lernen (ML), 2022.

ISO/IEC DIS 25059:2022-07 – Entwurf, System- und Software-Engineering – Qualitätskriterien und Bewertung von Systemen und Softwareprodukten (SquaRE) – Qualitätsmodell für KI-System.

ISO/IEC DIS 42001, Information Technology – Artificial intelligence – Management system.

ISO/IEC TR 24028:2020, Information technology – Artificial intelligence – Overview of trustworthiness in artificial intelligence.

ISO/IEC 5259 (alle Teile), Artificial intelligence – Data quality for analytics and machine learning (ML).

Kahneman, Daniel, Paul Slovic, and Amos Tversky. 1982 (1974). »Intuitive prediction: Biases and corrective procedures.« In: *Judgment under Uncertainty:*

Heuristics and Biases, edited by Kahneman, Daniel, Paul Slovic, and Amos Tversky. Cambridge: Cambridge University Press.

Moray, Neville. 2001. *Human and machines. Allocation of functions*. In: *People in Control: Human Factors in Control Room Design*, edited by Noyes, Janet M, and Matthew Bransby. London: Institution Of Electrical Engineers, 101–115.

Muhammad, Selma. 2022. *The Fairness Handbook*, Accessed April 1, 2023. https://www.amsterdamintelligence.com/resources/the-fairness-handbook.

Pentenrieder, Annelie and Jutta Weber. 2020. »Lucy Suchman.« In: *Technikanthropologie Handbuch Für Wissenschaft Und Studium*, edited by Heßler, Martina, Kevin Liggieri, and Nomos Verlagsgesellschaft. Baden-Baden Nomos, Edition Sigma, 215–224.

Puntschuh, Michael and Lajla Fetic. 2020. *Handreichung für die digitale Verwaltung. Algorithmische Assistenzsysteme gemeinwohlorientiert gestalten*. Bertelsmann Stiftung, Gütersloh. DOI: https://doi.org/10.11586/2020060.

Raisch, Sebastian, and Sebastian Krakowski. 2020. »Artificial Intelligence and Management: The Automation-Augmentation Paradox.« *Academy of Management Review*, February. DOI: https://doi.org/10.5465/2018.0072.

Rohde, Friederike et al. 2021. *Nachhaltigkeitskriterien für künstliche Intelligenz – Entwicklung eines Kriterien- und Indikatorensets für die Nachhaltigkeitsbewertung von KI-Systemen entlang des Lebenszyklus*. Accessed April 1, 2023. https://www.ioew.de/publikation/nachhaltigkeitskriterien_fuer_kuenstliche_intelligenz.

Schlick, Christopher, Ralph Bruder and Holger Luczak. 2010. *Arbeitswissenschaft*. 3. Auflage, Springer-Verlag Berlin Heidelberg.

Simonsen, Jesper, and Toni Robertson. 2013. *Routledge International Handbook of Participatory Design*. New York: Routledge.

Stowasser, Sascha and Oliver Suchy et al. (eds.). 2020. *Einführung von KI-Systemen in Unternehmen. Gestaltungsansätze für das Change-Management*. Whitepaper aus der Plattform Lernende Systeme, München.

Suchman, Lucille A. 2009. *Human-Machine Reconfigurations: Plans and Situated Actions*. Cambridge: Cambridge Univ. Pr.

Ulich, Eberhard. 2011. *Arbeitspsychologie*. Stuttgart.

Ulich, Eberhard. 2013. »Arbeitssysteme als Soziotechnische Systeme – eine Erinnerung.« *Journal Psychologie des Alltagshandelns/Psychology of Everyday Activity*, 6 (1), 4–12.

VDE, Bertelsmann Stiftung (Hg.). *From Principles to Practice – An interdisciplinary framework to operationalise AI ethics*. AI Ethics Impact Group, VDE Association for Electrical Electronic & Information Technologies e. V., Bertels-

mann Stiftung, 1–56, 2020. Accessed April 1, 2023. DOI: https://doi.org/10.11586/2020013.

VDI/VDE-MT 7100 – Entwurf, Lernförderliche Arbeitsgestaltung – Ziele, Nutzen, Begriffe.

VDI/VDE 3850–1:2014, Gebrauchstaugliche Gestaltung von Benutzungsschnittstellen für technische Anlagen – Konzepte, Prinzipien und grundsätzliche Empfehlungen.

VDE SPEC 90012:2022, VCIO based description of systems for AI trustworthiness characterization.

Zweig, Katharina A, Tobias Krafft, and Enno Park. 2021. *Sozioinformatik. Ein neuer Blick auf Informatik und Gesellschaft*. München.

Vertrauen im Kontext – Messung und Operationalisierung

Susanna Wolf, Jan Fiete Schütte, Marc Hauer, Christopher Koska

Einleitung

Vertrauen ist für den Erfolg von datengetriebenen Unternehmungen ein entscheidender Faktor. Vor allem mit Blick auf die radikalen Umbrüche, die durch die Einführung von neuen Technologien ausgelöst werden, gilt Vertrauen für viele Unternehmen als die wichtigste Ressource (Benrath 2019). Insbesondere wenn personenbezogene Daten erhoben und verarbeitet werden, müssen User z.B. darauf vertrauen können, dass ihre Daten sicher sind und alle Beteiligten verantwortungsbewusst mit ihren Informationen umgehen. Wenn ein Unternehmen das Vertrauen verliert, kann das schwerwiegende Auswirkungen haben, beispielsweise den Verlust von Mitarbeitenden oder Kundschaft. Dies kann zu negativen Reputationseffekten führen, aber auch zu rechtlichen Konsequenzen und großen finanziellen Schäden. Positiv gewendet kann Vertrauen ein wichtiger Wettbewerbsvorteil sein. Wenn User die Wahl zwischen verschiedenen IT-Unternehmen haben, gibt es Tendenzen, dass sie in der Regel demjenigen gegenüber am loyalsten sind, dem sie am meisten vertrauen (Capgemini 2017). Deshalb ist es für Unternehmen unerlässlich, Vertrauen in ihre digitalen Produkte und Dienstleistungen aufzubauen und aufrechtzuerhalten. Doch was ist eigentlich Vertrauen? Und wie lässt es sich messen und operationalisieren?

Ein Blick auf die Vertrauensforschung zeigt: Es gibt kein allgemeingültiges Erfolgsrezept, um Vertrauen in IT-Unternehmen aufzubauen oder zu festigen. Vertrauen ist graduell und entzieht sich einer eindimensionalen Messung. Der Akt des Vertrauensschenkens hängt u.a. von der subjektiven Bereitschaft ab, sowie von der konkreten Situation, in der sich diese Personen oder Personengruppen zu einem bestimmten Zeitpunkt befinden, aber auch von der spe-

zifischen technologischen Anwendungspraxis. Deshalb lassen sich die Messgrößen von Vertrauen rein als Variablen im Rahmen eines mehrdimensionalen Vertrauensmodells kategorisieren. Die Metriken, die für die Messung und Operationalisierung von Vertrauen in datengetriebenen Projekten verwendet werden, lassen sich nur über den jeweils konkreten Anwendungskontext bestimmen.

Dieser Beitrag stellt ein Konzept zur kontextspezifischen Operationalisierung und Messung von Vertrauen für datengetriebene Geschäftsmodelle vor. Das Konzept baut auf den Erkenntnissen und Erfahrungen des interdisziplinär agierenden *Datenethik-Hubs* auf, der initial als *Community of Practice* organisiert und von DATEV zur Umsetzung der unternehmensinternen *Directive Datenethik* (DATEV 2021) ins Leben gerufen wurde (PWC 2023).[1] Im ersten Schritt wird die begriffliche Grundlage für das hier vorgeschlagene Konzept entwickelt. Anschließend wird in einem zweiten Schritt ein Werkzeug vorgestellt, welches für vorhandene Datenbestände u.a. Verantwortlichkeiten und zulässige Verarbeitungszwecke katalogisieren kann und sich bei DATEV als eine zusätzliche Option für die verantwortungsbewusste Operationalisierung einer vertrauenswürdigen Datennutzungspraxis etabliert. Im dritten Schritt werden verschiedene Ansätze und Methoden zur Messung von Vertrauenswürdigkeit vorgestellt. Der vierte Schritt beinhaltet die konzeptionelle Erweiterung des Datenkatalog-Modells, um den Boden für die kontextspezifische Messung von Vertrauen vorzubereiten. Eine abschließende Reflexion fasst die Ergebnisse zusammen. Ziel des Artikels ist es, in den akademischen Diskurs zur Operationalisierung und Messung von Vertrauen einen praxisorientierten Beitrag einzubringen, indem er abstrakte Vertrauenskonzeptionen kritisch hinterfragt und anwendungsbezogen erweitert.

1 Mit dem Projekt »Digitale Verantwortung leben. Datenethik bei DATEV« hat das Softwareunternehmen 2021 den CDR Award (Corporate Digital Responsibility Award), verliehen von BVDW und Bayern Innovativ, in der Kategorie »CDR und Mitarbeitende« gewonnen. Ausgezeichnet wurde das Projekt, da die Directive Datenethik (»DATEV Datenethik-Leitlinie«) im Rahmen eines partizipativen Prozesses gemeinsam mit diversen Stakeholdern und Mitarbeitenden entwickelt und anschließend in einer initialen, bereichsübergreifenden Community of Practice, dem sogenannten Datenethik Hub, nachhaltig in der Organisationsstruktur verankert wurde.

Begriffliche Eingrenzung des Untersuchungsgegenstandes

Die sozialwissenschaftliche Organisationsforschung unterscheidet drei Formen von Vertrauen anhand eines Ebenenmodells (Wagenblass 2018, 1803):

- Auf der Makroebene wird das Vertrauen in gesamtgesellschaftliche Veränderungsprozesse diskutiert, also z.B. ganz grundsätzlich das Vertrauen in digitale Transformation, in Künstliche Intelligenz oder in Technik. Auch wenn wir im politischen Kontext vom Vertrauen in Demokratie oder im wirtschaftlichen Kontext vom Vertrauen in den materiellen Wert von Geld sprechen, dann befinden wir uns, wie Luhmann (1968) gezeigt hat, auf der Ebene des generalisierten Vertrauens oder Systemvertrauens.
- Auf der Mesoebene wird das Vertrauen in konkrete Institutionen oder projektbezogene Kooperationen adressiert. Die Vertrauenswürdigkeit von bestimmten Unternehmen, Institutionen und organisationsübergreifenden Kooperationen beruht hier insbesondere auf institutionalisierten Prozessen und Strukturen. Auf dieser Ebene geht es folglich um ein sehr spezifisches Vertrauen.
- Auf der Mikroebene wird das Vertrauen als ein Moment diskutiert, der aus der Interaktion und Beziehung zwischen den Vertreter*innen einer bestimmten Institution und den Vertrauensgebenden entstehen, beispielsweise zwischen den Mitarbeitenden eines Unternehmens und der Kundschaft eines Unternehmens. Auf dieser Ebene geht es vor allem darum, persönliches Vertrauen aufzubauen.

Im Fokus dieses Beitrags steht das spezifische Vertrauen, beispielsweise in eine konkrete Institution, also die Mesoebene. Darüber hinaus möchten wir zeigen, inwiefern das Vorgehen hierbei (insbesondere durch die Integration in eine CDR-Strategie)[2] auch positiv auf das generalisierte Vertrauen bzw. Systemvertrauen (Makroebene) einzahlt und von persönlichem Vertrauen (Mikroebene) getragen wird. Für die weitere Eingrenzung des Untersuchungsgegenstandes sind ferner analytische Vertrauensmodelle hilfreich, die zwischen vertrauensgebender Instanz (A), Vertrauensobjekt (X) und vertrauensnehmender

2 Einen wissenschaftlichen Einstieg in das Konzept der Corporate Digital Responsibility bietet u.a. Filipović (2023).

Instanz (B) unterscheiden (Michalski 2019, 44).[3] In unserem Fall ist die vertrauensnehmende Instanz (B) ein datenverarbeitendes Unternehmen. Für dieses Unternehmen ist es in einem nächsten Schritt hilfreich – mit Blick auf das Vertrauensobjekt (X) – zwischen der externen und internen Wert-Ebene von Vertrauen zu differenzieren (van de Poel 2013, 135ff):

- Intern: Aspekte wie Sicherheit, Verlässlichkeit, Robustheit, Wartungsfreundlichkeit, Kompatibilität, Qualität, Effektivität und Effizienz lassen sich beispielsweise mit Blick auf die Leistungsstärke als Eigenschaften (interne Werte) von technischen Artefakten (Vertrauensobjekten)[4] verstehen; insofern als Bedingungswerte für Vertrauen. In der Praxis werden die internen Eigenschaften der Vertrauensobjekte häufig als terminale Werte betrachtet. Tatsächlich handelt es sich dabei aber zumeist um instrumentelle Werte, die sich als Mittel zum Erreichen eines terminalen Wertes einsetzen lassen (van de Poel 2013, 137).
- Extern: Da es ganz wesentlich von der Bereitschaft der vertrauensgebenden Instanz (A) abhängt, dem technischen Bezugsobjekt (X) Vertrauen zu schenken, ist Vertrauen oder die Vertrauenswürdigkeit von »X« kein interner, sondern ein externer Wert. Denn die Bereitschaft Vertrauen zu schenken ist u.a. von den individuellen Einstellungen und persönlichen Präferenzen der vertrauensgebenden Instanz, aber auch von der konkreten Situation bzw. dem spezifischen Anwendungskontext abhängig.

Weiterhin findet in Schritt 4, »Kontextspezifische Erweiterung des Datenkatalog«, das ABI-Modell nach Schoorman et al. (2007) Anwendung, um konkrete Operationalisierungsmaßnahmen einzelnen Vertrauensaspekten zuzuordnen. Nach dem ABI-Modell sind diese *Ability* (Fähigkeit), *Benevolence* (Wohlwollen) und *Integrity* (Integrität). Durch diese Einstufung wird gezeigt, wie und warum die einzelnen Maßnahmen das Vertrauen im Sinne des oben genannten Ebenenmodells stärken.

Die Herausforderung bei der Messung und Operationalisierung von Vertrauen wird vor dem Hintergrund der hier skizzierten Modelle deutlich sicht-

3 Zum Entscheidungsvertrauen der vertrauensnehmenden Instanz vgl. auch Kaminski 2017, 167–169

4 Mit Blick auf den IEEE Std 7000™-2021 (IEEE Standard Model Process for Addressing Ethical Concerns during System Design) lassen sich die technischen Artefakte (Vertrauensobjekte) u.a. als SOI (system of interest) definieren.

bar. Vertrauen ist ein mehrdimensionaler Relationsbegriff. Die Frage »*Wer (A) vertraut wem (B) in Bezug auf was (X)?*« grenzt den Untersuchungsgegenstand dabei kontextbezogen ein. Da Vertrauen stets in spezifischer Referenz auf einen bestimmten Zusammenhang entsteht, zielt dieser Beitrag darauf ab, die Vertrauenswürdigkeit von konkreten Datenverarbeitungsszenarien (X) zu messen. Wird Vertrauen zu unspezifisch erfasst, entstehen zahlreiche Fehlerdimensionen. Deshalb ist der jeweils konkrete Bezug auf das Vertrauensobjekt (über seine effiziente Nutzbarkeit hinaus) ebenso relevant wie der auf die vertrauensnehmende Instanz (auch unabhängig vom Vertrauensobjekt). Nur auf diese Weise lassen sich Fehlerquellen durch eine einseitige Referenz identifizieren und eliminieren.

Anwendungsszenario: Operationalisierung einer vertrauenswürdigen Datennutzungspraxis

Datenverfügbarkeit gehört zu den Schlüsselkonzepten, wenn es um datengetriebene Geschäftsmodelle und Prozessautomatisierung geht. Um dies im Dienst der Genossenschaft voranzutreiben, beschäftigt sich das Data Office als Organisationseinheit bei DATEV mit Data Governance, also mit dem kontrollierten Verfügbarmachen von Daten durch adäquate Systeme und Prozesse sowie Daten zu erheben, zu verwalten und zu nutzen (Al-Ruithe 2019, Janssen 2020). Data Governance unterstützt dabei, Daten rechtskonform, effizient und wertorientiert zu verarbeiten. Der Anspruch für den hier angeführten Use Case ist auch, eine verantwortungsvolle Datenverarbeitungspraxis auf Basis der Unternehmenswerte von DATEV aufrechtzuerhalten. Das hier behandelte Anwendungsszenario zeigt exemplarisch, wie Unternehmenswerte – mitunter Vertrauenswürdigkeit – über ihre Verankerung in interne Prozesse fortlaufend in ihrer Umsetzungspraxis gestärkt werden. Das Beispiel geht insbesondere auf zwei Aspekte ein: zum einen wie digitale Verantwortung (im Sinne von Corporate Digital Responsibility, CDR) bei DATEV ausgerichtet ist, zum anderen wie sie exemplarisch zum Tragen kommt. Der hier gewählte Fokus bezieht sich auf das konzeptionelle Entwickeln und Etablieren eines wertorientierten Datenkatalogs.

Mit Bezug auf die unternehmerische Verantwortung (Corporate Social Responsibility, CSR) und ihrer Vertiefungsrichtung CDR (Anzinger 2021, Seidl 2022) können Unternehmen als Vertrauensnehmer*innen sicherstellen, dass ihr Wertbezug – somit grundsätzlich auch ihr Bezug auf Vertrauens-

würdigkeit – beim Entwickeln und Betreiben von Technologie gegenüber ihren Stakeholdern als Vertrauensgeber*innen präsent bleibt und dadurch mittelbar wie unmittelbar in Produkten, Systemen und Services als Vertrauensobjekte zur Darstellung kommt. CSR bei DATEV ist nach der klassischen Triple-Bottom-Line gegliedert: in die ökonomische, ökologische und soziale Säule (Slaper und Hall 2011). Das bedeutet: DATEV agiert im Bewusstsein der Verantwortung für Wirtschaft, Umwelt und Gesellschaft (DATEV 2023a). Digitale Verantwortung verortet DATEV als Softwarehaus bei nachhaltiger Innovation und vor diesem Hintergrund in der ökonomischen Säule. Insofern bewegt sich das hier beleuchtete Beispiel auf der Mesoebene und adressiert insbesondere wertorientierte Data Governance als Voraussetzung für die Leistungsstärke von datengetriebenen Produkten (internes Vertrauen) sowie deren vertrauenswürdiges Potential aus Stakeholder-Perspektive (externes Vertrauen). Dies zeigt sich auch mit Blick auf die Operationalisierung der erwähnten Unternehmenswerte.

Sie strukturieren dabei die Fokuspunkte von Digitaler Verantwortung im Rahmen von DATEVs Datenverarbeitungspraxis, festgehalten in der Directive Datenethik. Diese Directive als interne Regulatorik ist direkt im DATEV-Verhaltenskodex – *Code of Business Conduct* (CoBC) (ebd.) verankert und zeigt durch den konsequenten Rückbezug: Digitale Verantwortung leitet sich aus den grundsätzlichen Unternehmenswerten von DATEV ab. *Vertrauenswürdig*, *Leistungsstark*, *Partnerschaftlich*, *Führend* und *Nachhaltig* sind die Unternehmenswerte, die hierbei zum Tragen kommen:

- Vertrauenswürdigkeit ist dabei insofern kontextgebunden, als sie darauf abzielt, gemeinsam mit den je beteiligten Stakeholdern Chancen und Herausforderungen der entsprechenden datenbasierten Lösungen für den individuellen Fall angemessen zu reflektieren, »um so ein gemeinsames Verständnis für langfristige Ziele und nachhaltiges Handeln zu entwickeln« (DATEV 2021).
- Dieser Anspruch findet sich im »ganzheitlichen Wertemanagement« (ebd.) bei Leistungsstark wieder. Hier geht es darum, Innovationen wertorientiert voranzutreiben und mit Blick auf Technologie-Entwicklung und -Einsatz sicherzustellen: »Im Zentrum steht dabei immer der Mensch mit seinen individuellen Bedürfnissen und Interessen« (ebd.). Das bedeutet auch, digitale Lösungen zu entwickeln, die Mitglieder und Nutzer*innen bei ihrer Arbeit bestmöglich unterstützen und gut zu ihnen passen. Wenn

es darum geht, mit datenbasierten Innovationen Stakeholdergruppen gerecht zu werden, impliziert dies auch Fairness.
- Der Wert Partnerschaftlich geht insbesondere auf Mitgestaltungsoptionen ein: über aktiven Stakeholder-Einbezug zur Entwicklung und Weiterentwicklung von DATEV-Produkten und -Services sowie Feedbackschleifen geht es darum, »Datensouveränität genossenschaftlich [zu] ermöglichen« (ebd.).
- Um »das Potential für einen zukunftsweisenden Umgang mit Daten zu erkennen und verantwortungsvoll auszuschöpfen« (ebd.), fördert DATEV im Sinne des Wertes Führend das Bewusstsein der Mitarbeitenden für wertorientierte Technologiegestaltung und beteiligt sich am gesellschaftlichen Diskurs.
- Dabei stellt die Directive Datenethik final klar: die Unternehmenswerte bleiben in ihrem Bezug auf eine verantwortungsvolle Datenverarbeitungspraxis nicht dadurch beständig und relevant, dass sie einmalig regulatorisch fixiert sind. Es gilt, sie fortlaufend diskursiv zu prüfen, umzusetzen und darüber zu aktualisieren. Das bedeutet, sie lebendig werden zu lassen und zu leben – unter anderem über fortlaufendes Use-Case-bezogenes Umsetzen. Diesen Aspekt hält die Directive mit Bezug auf einen fünften Wert fest: »Nachhaltig. Gemeinsam Standards leben« (ebd.).

Wie die Unternehmenswerte von DATEV bei der Datenverarbeitungspraxis verinnerlicht und gleichzeitig aktuell bleiben, daran hat verantwortungsvolle Data Governance grundsätzlich ihren Anteil. Data Governance bedeutet hier unter anderem: Datenverarbeitungsprozesse auch infrastrukturell und organisatorisch zu unterstützen, um die Daten-Wertschöpfung innerhalb eines definierten regulatorischen sowie wertbezogenen Rahmens effizient zu gestalten. DATEV zielt darauf ab, verantwortungsvolle Data Governance dabei als sogenannte *Good Data Governance* zu operationalisieren: Neben Effizienz und Wertschöpfung ist die Wertperspektive im Sinne Digitaler Verantwortung (CDR) hier ebenso im Fokus wie partnerschaftliche Vernetzung mit Bezug auf vertrauenswürdige Datenräume insbesondere im europäischen Kontext. Insofern lässt sich auch dieser Anwendungsfall mit Bezug auf den hier thematisierten wertorientierten fachlichen Datenkatalog als ein Werkzeug von *Good Data Governance* betrachten, das dem zuletzt genannten Umsetzungsanspruch der Directive Datenethik Rechnung trägt.

Klar ist, dass es sich hierbei zunächst um interne Prozesse handelt. Der Wertbezug mit Blick auf die vertrauensgebende Person liegt nun darin, dass

die fortwährende prozessuale Referenz auf Unternehmenswerte – auch hinsichtlich perspektivischer Datenverarbeitungsvorhaben – dem Anspruch Rechnung trägt, dass die Entwicklung und der Betrieb von Technologie nicht losgelöst von CDR-Aspekten gedacht und umgesetzt werden. So sind interne (für diesen Fall: leistungsstarke datengetriebene Produkte) wie externe Vertrauensdimension (für diesen Fall: Vertrauen externer Stakeholder) stets reflektiert. Im genannten Wert-Kontext zueinander stärken sie über interne Prozesse die Werteverbindung von (zukünftigen) Produkten, Systemen und Services. Der Wertbezug kommt schließlich während der Nutzung durch entsprechende Stakeholder faktisch zum Tragen.

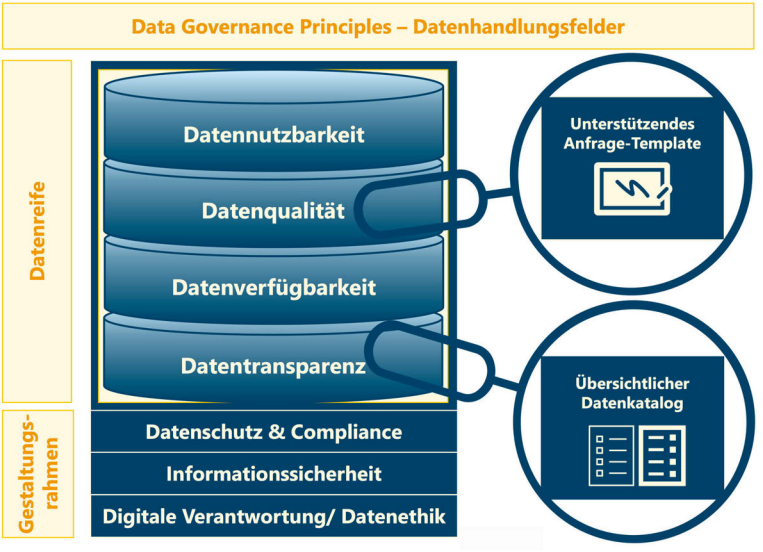


Abbildung 1: Data Governance Principles – Datenhandlungsfelder

Wie Mitarbeitende bei DATEV Datenkultur gemeinsam gestalten, halten die *Data Principles* fest. Sie beschreiben wesentliche Handlungsfelder im Rahmen der Datenverarbeitungspraxis. Dass sich das Nutzbarmachen von Daten an grundlegenden strategie- und wertorientierten Zielen nachhaltig ausrichtet, ist ein Grund dafür, dass sie zentral im *Data Office* festgehalten sind und dort – gemessen an je verschiedenen Kriterien – gemonitort werden. Einzelne Um-

setzungsmethoden und -werkzeuge, sogenannte *Wegbereiter*, operationalisieren die *Data Principles*. Die Abbildung zeigt, dass es zwei verschiedene Arten von Data Principles gibt:

1) Gestaltungsprinzipien als regulatorische und Werte-Basis legen die Leitplanken bei der Nutzbarmachung von Daten fest. Zu den Gestaltungsprinzipien gehören neben Datenschutz, Informationssicherheit und Compliance auch Digitale Verantwortung und Datenethik. Das bedeutet: die oben genannten Unternehmenswerte für eine verantwortungsvolle Datenverarbeitungspraxis sind über den Verweis auf die Directive Datenethik als Leitplanke direkt in die Data Principles integriert. Diese Verknüpfung bildet die normative Basis für wertorientierte Data Governance.
2) Datenreifeprinzipien treiben innerhalb dieses Gestaltungsrahmens den Reifegrad des Datenbestands voran – mit Fokus auf Transparenz, Verfügbarkeit, Qualität und Nutzbarkeit.

Abbildung 1 zeigt zwei Wegbereiter, die in den Datenreifeprinzipien Transparenz und Qualität verankert sind: den fachlichen Datenkatalog, sowie ein unterstützendes Anfrage-Template für Mitarbeitende, mit Datenverarbeitungsvorhaben, beispielsweise zur Produktentwicklung. Diese beiden Wegbereiter mit dem jeweiligen Prinzip, dem sie zur Umsetzung verhelfen, sollen im Rahmen des Anwendungskontextes in ihrer Wertorientierung näher betrachtet und daraufhin auf die unten dargestellte *Data Prosumer Journey* bezogen werden.

Die Journey dient in ihrer grundsätzlichen Form dazu, einen kontrollierten Lebenszyklus von Datenprodukten in Kombination mit verbundenen Bedarfs- und Umsetzungsprozessen der verantwortlichen und beteiligten Mitarbeitenden zu ermöglichen. Für datengetriebene Projekte sind das beispielsweise die verantwortlichen Mitarbeitenden, die eine bestimmte Datenanalyse umsetzen möchten, wie sogenannte Data Scientists. Ein Datenprodukt kann also beispielsweise eine Datenanalyse sein, die auswertet, wie effizient eine Software interne Prozesse von Kunden*innen automatisiert. Der genannte Lebenszyklus umfasst die verschiedenen Phasen des Datenprodukts – von den Konzeptions- bis zu den Löschvorgängen.

Abbildung 2 zeigt, inwieweit bei DATEV der Data Scientist als Data Prosumer im Fokus steht. *Prosumer* setzt sich dabei begrifflich aus *Consumer* und *Producer* zusammen, was bedeutet: Data Scientists als *Prosumer* fragen für die Analyse Daten an (*Consumer-Rolle*) und stellen in Folge über ihre Analy-

seergebnisse auch wieder Daten (als Datenprodukte) bereit (*Producer-Rolle*). Mit Fokus auf die *Data Principles* tritt an dieser Stelle Datentransparenz in den Vordergrund: für die Rolle *Data Consumer* ist wichtig, dass Daten für die geplante Analyse leicht im Bestand auffindbar sind, dass sie auf Metadatenbasis in ihren Eigenschaften (bspw. Vertraulichkeit oder zweckgebundene Verarbeitung) verständlich sind, sowie, dass leicht zuordenbar ist, wofür sie sinnvoll und wertorientiert genutzt werden können und wofür nicht. Diese Anforderungen spielen für *Producer* ebenso eine wichtige Rolle: nur wenn die Aspekte von Auffindbarkeit, Verständlichkeit und Zuordenbarkeit für sie transparent sind, gelingt es, entstehende Analysedaten wertschöpfend in den Bestand zu integrieren, um diese im Rahmen ihrer Verarbeitungszwecke wieder zur Verfügung zu stellen.

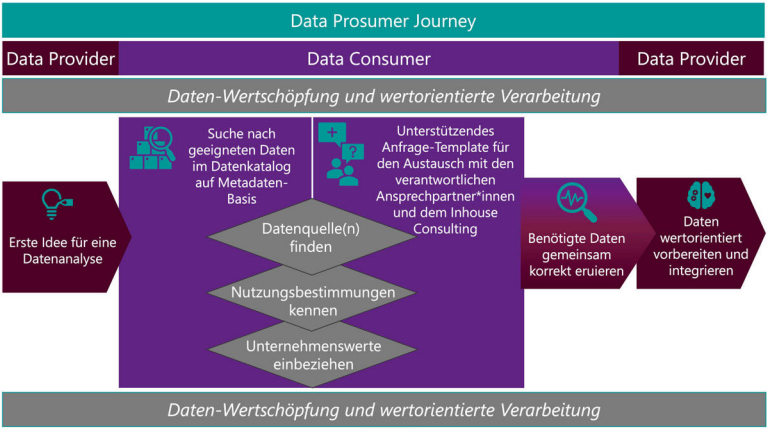


Abbildung 2: Ausschnitt aus der Data Prosumer Journey

Voraussetzung für diese gute Übersicht und damit das Datenreifeprinzip Transparenz bietet ein fachlicher Datenkatalog. Damit ist die fachliche Katalogisierung vorhandener Datenbestände auf Metadatenbasis gemeint. Verknüpft sind die Datensätze unter anderem mit entsprechenden Informationen über zulässige Verarbeitungszwecke sowie verantwortliche Ansprechpartner*innen. Dadurch ist ein zuverlässiges Wissens- und Informationsmanagement gegeben, um die Daten im Dienst der Genossenschaft effizient und vertrauenswürdig verfügbar zu machen. Hierbei zeigt sich,

dass der fachliche Datenkatalog als Werkzeug Unternehmen eine verantwortungsvolle Data Governance ermöglichen, CDR-Orientierung unterstützen und so Unternehmen in ihrer Rolle als vertrauensnehmende Instanz stärken kann. Er dient dazu, Leistungsstärke von Datenprodukten zu unterstützen (internes Vertrauen) und dadurch letztlich dazu, darauf einzuzahlen, wie dieses Potential für Vertrauenswürdigkeit wahrgenommen wird (externes Vertrauen).

Vor diesem Hintergrund benötigen *Data Prosumer* Fachkenntnisse zur unmissverständlichen Anwendung des Datenkatalogs.

Hierzu gehören:

- Verständnis der mit den Datensätzen verbundenen Logik für eine rechtskonforme Verarbeitung
- Kenntnis der Verantwortlichkeiten und Prozesse
- Transfer der Unternehmenswerte auf eine wertorientierte Datenverarbeitungspraxis
- Neben diesem Wissen ist das stete Bewusstsein für wertorientierte Verarbeitung relevant. Dazu gehört auch, Herausforderungen bei Datenanalysen wie einer ungewollten Verzerrung der Analyseergebnisse stets im Blick zu behalten. Über Dialog- und Weiterbildungsformate ist dabei relevant, folgende Aspekte zu adressieren:
  - Bewusstsein für relevante Verzerrungs-Arten (Bias)
  - Kenntnis von Optionen, diese Risiken zu reduzieren
  - Wissen um die Optionen, gewollten Bias (bspw. positive Diskriminierung, um Fairness-Potenziale auszuschöpfen) zu integrieren

Das Wissen um sowie die Sensibilität hinsichtlich der oben genannten Punkte zahlt auf ein weiteres *Data Principle* ein: Datenqualität. Stichworte in diesem Zusammenhang sind unter anderem Korrektheit, Aktualität und Redundanzfreiheit der Daten. In Prüfung ist vor diesem Hintergrund auch ein digitales Template für Anfragen mit erster und zweiter Ebene. Das Template integriert auf der ersten Ebene grundsätzliche fachliche und organisatorische Informationen, wie zulässige Verarbeitungszwecke und verantwortliche Ansprechpartner*innen, eine Skizze der geplanten Datenverarbeitung, auf welche Unternehmenswerte das Vorhaben grundsätzlich einzahlt und welche Stakeholder insbesondere zu den betroffenen Zielgruppen gehören. Die zweite Ebene zielt darauf ab, Maßnahmen, Chancen und Herausforderungen der Da-

tenverarbeitung – direkt unter Berücksichtigung von Wertereferenzen – für das Unternehmen und die jeweilige Zielgruppe im Voraus zu bedenken.

Instrumente zur Messung von Vertrauenswürdigkeit

Um die voran beschriebene Operationalisierung durchzuführen und ihren Erfolg zu bewerten, ist es erforderlich, konkrete Messinstrumente für Vertrauen zu finden und anzuwenden. Eine Messung des Vertrauens in ein Produkt (wir folgen hier dem Normverständnis, dass auch Dienstleistungen Produkte sind) wäre prinzipiell über die Akzeptanzrate, oder die Akzeptanzgeschwindigkeit denkbar. Jedoch gibt es hierbei starke Einschränkungen der Aussagekraft. Zum einen gibt es keine offiziell genormten Richtwerte, ab denen man von einem hohen oder niedrigen Vertrauen sprechen könnte. Darüber hinaus können die Gründe für die (Nicht-)Akzeptanz in einer Vielzahl an möglichen Ursachen und deren Kombination miteinander – Abseits der Vertrauensmessung – begründet liegen. Es kann beispielsweise daran liegen, dass es keine oder nur schlechtere/teurere Alternativen gibt oder, dass das Produkt oder die Dienstleistung ein Novum darstellt, für das zunächst einmal ein Markt entstehen muss. Deshalb ist es, wie oben skizziert, sinnvoller, die Vertrauenswürdigkeit zu messen und diese Messung an die Stakeholder zu kommunizieren, um damit die Bildung von Vertrauen zu unterstützen.

Eine gängige Möglichkeit die Vertrauenswürdigkeit in ein Produkt oder eine Dienstleistung zu bewerten, ist das Durchführen und Auswerten von Tests. Der Testbegriff ist sehr umfassend und nicht einheitlich definiert. Gemäß ISO 29119–1:2022[5] ist Testen ein »*Satz von Aktivitäten, die durchgeführt werden, um die Entdeckung und/oder Bewertung von Eigenschaften eines oder mehrerer Testelemente zu erleichtern*«, während ein Testelement »das Ergebnis einer Aktivität ist, wie z.B. Management, Entwicklung, Wartung, Tests selbst oder andere unterstützende Prozesse«. Für einen fachlichen Datenkatalog bietet es sich u.a. an, Tests bezüglich der Datenqualität (sind die Informationen bspw. vollständig, ausreichend aktuell, korrekt, redundanzfrei) durchzuführen. Für Anwendungen auf Basis einer initialen Metadaten-Recherche spielen Tests zu den Themen Zweckbindung, Vorhersagequalität (bspw. bei Analysen) und Nicht-Diskriminierung potenziell eine große Rolle. Derartige Tests basieren in der Regel auf der Berechnung von sogenannten Qualitäts- und Fairnessmaßen (Verma und

5 Software- und Systemengineering – Software-Test- Part 1: Allgemeine Konzepte.

Rubin 2018). Erreichen die gewählten Maße einen vorab definierten Mindestwert, gilt der entsprechende Test als bestanden. Sollte das Resultat eines Tests negativ ausfallen, wird das Produkt entsprechend den erwarteten Eigenschaften verbessert. Je besser also Produkte, bzw. die einzelnen Komponenten eines Produktes, getestet sind, desto höher fällt ihre Vertrauenswürdigkeit bezüglich ihrer Eigenschaften und Leistungen aus.

Tests allein sind dabei nicht ausreichend, um pauschal eine hohe Vertrauenswürdigkeit zu attestieren. Insbesondere, wenn Ziel- oder Interessenskonflikte vorliegen, zum Beispiel, weil in datengetriebenen Projekten keine klaren Rollen- und Funktionstrennungen existieren (wie beispielsweise Planung und Umsetzung durch die gleichen Personen), könnten Tests gezielt nur für Aspekte konzipiert werden, die unabhängig davon umgesetzt werden können, um dies als Qualitätssicherung zu kommunizieren und auf dieser Basis eine hohe Systemqualität zu behaupten. Insofern ist es empfehlenswert, eine Rollentrennung zu berücksichtigen zwischen denjenigen Expert*innen, die Produkte konzipieren und denjenigen, die Tests als kontrollierende Instanz in die Qualitätssicherung integrieren.

Dabei bestehen – wie grundsätzlich bei der Durchführung von Analysen und der Kommunikation ihrer Ergebnisse – Risiken: bspw. dass Testergebnisse nicht adäquat interpretiert oder missverständlich kommuniziert werden.[6] Eine Option, diese Risiken zu reduzieren, kann in sogenannten Audits liegen. Mit Blick auf die Bedeutung von Audits durch unabhängige Prüfinstitutionen für Vertrauenswürdigkeit von Nutzenden ist erwähnenswert, dass diese neben der Selbstverpflichtung von Unternehmen im Rahmen von Ethikleitlinien eine bedeutende Rolle spielen.[7]

Auch für den Begriff Audit gibt es unterschiedliche Auffassungen. Zwei davon sind besonders relevant: Inspektionen auf der Grundlage von Normung (z.B. der Korpus der ISO-Normungsdokumente, der die Grundlage für akkreditierte Zertifizierungen bildet) und Bias-Audits von Plattformen und Instituten (u.a. Ada Lovelace Institute in Europe (2023) oder die Initiative Zertifizierte

6 Klar ist, dass es auch Vorfälle gibt, bei denen Ergebnisse wissentlich falsch kommuniziert werden. Dies bewegt sich allerdings insbesondere im Bereich dessen, was rechtlich unzulässig ist (legale Ebene) und nicht mehr rein im datenethischen Bereich (legitime Ebene).

7 Gerade für Technologien wie Künstliche Intelligenz ist dies der Fall (ZVKI 2022; TÜV-Verband 2021).

KI des Fraunhofer IAIS (2023)). Nach ISO 19011:2018 ist ein Audit ein »*systematischer, unabhängiger und dokumentierter Prozess zum Erlangen von objektiven Nachweisen [.] und deren objektiver Auswertung, um zu bestimmen, inwieweit die Auditkriterien [.] erfüllt sind*« (ISO 19011: 2018).

Diese Definition bildet die Grundlage für alle Verweise auf Audits im gesamten Korpus der ISO-Normungsdokumente. Sie unterscheidet ausdrücklich zwischen Erstparteien-Audit (*1st party audit*), Zweitparteien-Audit (*2nd party audit*) und Drittparteien-Audit (*3rd party audit*). Erstparteien-Audits werden von der Organisation selbst oder in ihrem Namen durchgeführt (ISO 19011: 2018). Sie sind geeignet, um interne Vertrauenswürdigkeit in die Funktionalität interner Regelungen und Prozesse sowie deren Gültigkeit aufzubauen. Es geht also primär um die Vertrauenswürdigkeit der internen Stakeholder, bspw. in die Wirksamkeit des eigenen Managementsystems. Zweitparteien-Audits werden auf Initiative von externen Stakeholdern durchgeführt, die ein Interesse an den Qualitätsstandards einer bestimmten Organisation haben (z.B. Kunden, oder Institutionen in deren Auftrag). Drittparteien-Audits werden von unabhängigen Prüforganisationen durchgeführt, die bspw. Konformitätszertifikate oder -registrierungen ausstellen, oder von staatlichen Stellen (ISO 19011: 2018). Bei Zweit- und Drittparteien-Audits handelt es sich also um Audits durch Externe. Das bedeutet, dass Prüfpersonal Zugang zu den relevanten Informationen und Systemen erhält, um bspw. ein Produkt oder einen Prozess zu untersuchen. Dies schließt auch Tests ein. Die Aussagekraft eines Audits hängt damit auch von der Vertrauenswürdigkeit des auditierenden Prüfpersonals und des durchgeführten Auditprozesses ab. Diese Vertrauenswürdigkeit wiederum kann durch Zertifikate, die die Prüfinstitutionen selbst innehaben bemessen werden.

Ein Zertifikat ist eine Auszeichnung durch eine kompetente Instanz, die einer Person oder Sache eine bestimmte Eigenschaft bescheinigt und in dem hier behandelten Kontext auf die Vertrauenswürdigkeit eines Vertrauensnehmers oder eines Vertrauensobjekts abstellt. Bei Zertifikaten ist grundsätzlich zwischen allgemeinen und akkreditierten Zertifikaten zu unterscheiden. Im Grunde kann jeder ein Zertifikat als Bescheinigung ausstellen. Die Aussagekraft eines Zertifikats durch eine Instanz, die sich (bspw. über entsprechende Nachweise) noch kein Vertrauen verdient hat, ist jedoch sehr gering. Je vertrauenswürdiger eine Instanz ist, desto aussagekräftiger sind Zertifikate, die diese Instanz ausstellt. Am vertrauenswürdigsten sind die sogenannten akkreditierten Zertifikate, die nur durch akkreditierte Zertifizierungsstellen ausgestellt werden können. Diese wiederum sind durch die Akkreditierungsstel-

le eines Landes legitimiert (Bundesnetzagentur 2018; Datenschutzkonferenz 2020). Der europäische Rechtsrahmen stellt sicher, dass es in jedem Land der Europäischen Union genau eine Akkreditierungsstelle gibt (in Deutschland ist das die Deutsche Akkreditierungsstelle GmbH, DakkS), die (auf der Makroebene) die Kompetenz der Zertifizierungsstellen (auf der Mesoebene) bestätigt. Da akkreditierte Zertifikate auf Normen basieren müssen, sind die Mindestanforderungen zur Ausstellung akkreditierter Zertifikate für Zertifizierungsunternehmen vielfach reflektiert und eindeutig formuliert. Akkreditierte Zertifikate können sich auf verschiedene Vertrauensobjekte der jeweiligen Zertifizierungsstelle beziehen u.a.: ihre Managementsysteme (ISO/IEC 17021–1: 2015), ihre Mitarbeitenden im Sinne einer Personenzertifizierung wie Sachverständige (ISO/IEC 17024: 2021) und ihre Produkte und Prozesse (ISO/IEC 17065: 2013; Datenschutzkonferenz 2019).

Es gibt noch über Zertifizierungen hinausgehende Konzepte zur Bemessung der Vertrauenswürdigkeit. Besonders relevant sind dabei Kriterien-basierte Ansätze. Dafür wird eine Sammlung an Kriterien erstellt, die als Indikatoren für Vertrauenswürdigkeit fungieren. Das Maß an Vertrauenswürdigkeit kann dann zum Beispiel über die Anzahl erfüllter Kriterien ausgedrückt werden. Da je nach Produkt nicht alle Kriterien anwendbar sind, könnte auch das Verhältnis zwischen erfüllten und nichterfüllten Kriterien ein angemessenes Maß darstellen. Ein solches Maß ist nicht dafür geeignet die absolute Vertrauenswürdigkeit auszudrücken, jedoch um die relative Vertrauenswürdigkeit zwischen ähnlichen Produkten und Dienstleistungen (basierend auf dem gleichen Kriterienkatalog)[8] zu ermitteln. Einen ähnlichen Ansatz verfolgen auch aktuelle Überlegungen zum Thema Corporate Digital Responsibility.[9] Darüber hinaus können Checklisten auch alle bisher genannten Aspekte (also Tests, Auditierung und Zertifizierung) umfassen und ermöglichen damit eine holistische Bemessung der Vertrauenswürdigkeit. Gleichzeitig gibt es keine korrekte oder vollständige Sammlung an Aspekten, die in einer Checkliste erfragt werden sollten. Somit muss eine geeignete Checkliste anwendungsbezogen erstellt, sowie vergleich- und überprüfbar sein, um belastbare Aussagen daraus ableiten zu können.

8 Angemessene Kriterien für einen solchen Kriterienkatalog können mit Blick auf KI zum Beispiel auf Basis des VCIO Modells (VDE 2022, 7–47) gewonnen werden.

9 Es gilt dabei auch, die verschiedenen Konzepte von Digitaler Ethik zu berücksichtigen (Koska 2023).

Parallel zu den genannten Möglichkeiten kann die Vertrauenswürdigkeit eines Produktes auch im Kontext einer Risikobemessung bzw. Technikfolgenabschätzung betrachtet werden. Ein höheres Risiko verweist auf eine erhöhte Verletzbarkeit und erfordert deshalb ein höheres Maß an Vertrauen (Koska et al. 2024). Dadurch entsteht ein Bedarf, Mechanismen einzusetzen, die das Risiko reduzieren, bzw. Konzepte, mit den Risikofolgen kontrolliert umzugehen (Risikomanagement), um somit die Vertrauenswürdigkeit zu stärken. Diesen Weg geht zum Beispiel auch der Entwurf des AI-Acts (European Commission 2021), der eine Risikoklassifizierung auf Basis der Technologie und des Anwendungskontextes [AI-Act Annex II, Annex III und Art. 52] vorsieht. Je nach Klassifizierung werden unterschiedliche Anforderungen an ein Produkt gestellt, zum Beispiel in Bezug auf Tests, Auditierung, Zertifizierung und Transparenz. Die Kopplung der obligatorischen Etablierung von Schutzmechanismen an eine vorige Risikobemessung bzw. Technikfolgenabschätzung ist ein häufig diskutierter Lösungsansatz, um Vertrauenswürdigkeit herzustellen, ohne eine Überregulierung zu verwirklichen (Hallensleben et al. 2020).

Kontextspezifische Erweiterung des Datenkatalogs

Der folgende Abschnitt erläutert insbesondere am Beispiel des Datenkatalogs, mit welchen Methoden und Werkzeugen bei der Operationalisierung vorgegangen werden kann sowie was es dabei zu beachten gilt. Ziel ist es, eine ganzheitliche Basis für Vertrauenswürdigkeit zu schaffen – nicht nur im Rahmen des Datenkatalogs selbst, sondern mit vorausschauendem Blick auf die damit verknüpften Prozesse und Produkte.

Hierfür lässt sich wertorientierte Technologiegestaltung bei datengetriebenen Projekten grundsätzlich in drei Phasen unterteilen (IEEE SA 2021):

1) Begutachtung der Ist-Lage des *System of Interest* und Entwicklung von wertorientierten Zielen
2) Maßnahmenplanung und -durchführung – *Ethical Value Requirements*
3) Maßnahmenbewertung und Iteration – Qualitätssicherung der *Ethical Value Requirements*

Für die Kategorisierung einzelner Maßnahmen bietet sich das ABI-Modell an. Es ist ein Konzept für die Modellierung von Vertrauen, beschrieben durch

Schoorman et al. (2007) auf Grundlage von McAllister (1995). Im Fokus stehen dabei folgende Vertrauensaspekte:

- Ability (Fähigkeit): Die Fähigkeit des Vertrauensnehmers, die erwarteten/angebotenen Aufgaben innerhalb der Vertrauensbeziehung zu erfüllen.
- Benevolence (guter Wille): Der gute Wille gegenüber dem Vertrauensgeber, dessen Interessen zu wahren.
- Integrity (Integrität): Ein schlüssiges und konsistentes Verhalten des Vertrauensnehmers u.a. durch die fortwährende Einhaltung (selbst)gesetzter Grundsätze, welche für den Vertrauensgeber Relevanz haben.

Durch die Zuordnung von Maßnahmen zu diesen Schwerpunkten, wird deutlich, welches Ziel eine Maßnahme verfolgt. Ebenso kann die Leistung des Vertrauensnehmers für diese drei Aspekte gemessen werden, beispielsweise durch Umfragen, und es können daraus nötige Maßnahmen abgeleitet werden.

Auf Prozessebene kann hierfür bspw. der Standard IEEE 7000–2021 (IEEE SA 2021) hinsichtlich vorgesehener Prüfschritte und -kriterien berücksichtigt werden, insofern Umfang und Inhalt zum Zielprodukt passen. Es handelt sich um eine Prozessnorm für Produktentwicklung unter Berücksichtigung einer Werte-Reflexion (*Ethics by Design*), welche für das gesamte Phasenmodell anwendbar ist.

Phase 1: Begutachtung der Ist-Lage des *System of Interest* und Entwicklung von wertorientierten Zielen

Zunächst gilt es, den Anwendungskontext des Vertrauensobjekts (*System of Interest, SOI [IEEE SA. 2021, 22]*) zu skizzieren. Um Vertrauenswürdigkeit für ein bestimmtes Produkt (auf der Mesoebene) aufzubauen, ist unter anderem die Identifizierung und der Einbezug der relevanten Stakeholdergruppen (bei Bedarf über entsprechende repräsentative Personen) entscheidend. Der Datenkatalogs als SOI integriert sich weiterhin in eine bestehende Produktlandschaft. Der Katalog bietet aber nicht nur die Basis dafür, eine Vielzahl an unterschiedlichen Datennutzungsszenarien zu verwalten, sondern ermöglicht auch neue Anwendungsfelder wertorientiert zu erschließen. In diesem Sinne kann der Datenkatalog u.a. als ein Management-Tool betrachtet werden, das dazu dient, datengetriebene Projekte vertrauenswürdig zu verwalten

und zu gestalten, insbesondere in Bezug auf deren Beziehung zueinander, wie sie in Systemen mit gegenseitiger Referenz auftreten können (*System of Systems*, SoS *[IEEE SA. 2021, 22]*).

Bei der Entwicklung von Zielen gilt es, bestehende Richtlinien, Kodizes und Werte innerhalb des Unternehmens anzuwenden. Dafür setzt DATEV unter anderem die *Directive Datenethik* ein. Unter Einsatz der fünf Unternehmenswerte werden Handlungsbedarfe ermittelt und kontextspezifische Ziele gesetzt sowie Kriterienkataloge erstellt. Durch die fortwährende Umsetzung der *Directive* zielt DATEV darauf ab, auf die Integrität nach dem ABI-Modell einzuzahlen. Der Datenkatalog ist ein Werkzeug, welches insbesondere durch geprüfte Daten die Intention unterstützt, dass die ihm zugrundeliegenden Werte in datengetriebenen Projekten berücksichtigt werden. Dies kann einerseits dazu beitragen, die Integrität (*Integrity*) zu verbessern und andererseits das Zutrauen in die Fähigkeiten (*Ability*) des Unternehmens stärken, da hochwertige Datensätze die Leistung der Produkte erhöhen.[10]

Außerdem können auch Werkzeuge aus der Organisationsmethodik genutzt werden, wie *Canvas*-Formate (*Ethics Canvas* (Online Ethics Canvas n. d.), *Data Ethics Canvas* (Open Data Institute n. d.), *Digital Product Ethics Canvas* (Gerlach 2019)), welche ebenfalls dabei unterstützen, die Rahmenbedingungen und Implikationen der Produkte strukturiert zu reflektieren und zu überblicken. Ein *Canvas* ist eine visuelle Organisationsmethodik, die komplexe Projekte übersichtlich und nachvollziehbar darstellt. Das Canvas bietet den Rahmen für mehrere untergeordnete Kategorien, die als Schlüsselelemente eines Projekts zu skizzieren sind. Im Rahmen des Datenkatalogs wird derzeit beispielsweise mit Canvas-Formaten gearbeitet, um den Anfrageprozess durch interne Stakeholder (bspw. o.g. Data Prosumer) zu strukturieren. Dies bietet Potenzial für einen effektiven Anfrage-, Beratungs- und Datenbereitstellungsprozess.

Für weitere Tätigkeiten, wie die der Technikfolgenabschätzung mit Fokus auf die Wertorientierung, gibt es eine große Auswahl von Werkzeugen. Als Einstieg bietet sich das Verzeichnis der OECD an, welches eine Vielzahl an Hilfsmitteln zum Thema »Vertrauenswürdige KI« umfasst und zudem umfangreiche Filter- und Suchoptionen bietet (OECD.AI Policy Observatory n.

10 Der erwähnte IEEE 7000 beginnt analog zu Phase 1 mit den Fragen nach dem Concept of Operations und dem Context Exploration Process (Kap. 7). Hierbei werden die Ziele der Organisation, sowie das ethische Umfeld ergründet. Im nächsten Schritt, dem Ethical Values Elicitation and Priorization Process (Kap. 8), werden die kontextspezifischen Werte ermittelt und priorisiert.

d.). Für die Verwendung in Projekten oder Unternehmen lässt sich empfehlen, eine eigene Werkzeugsammlung zu erstellen und zu pflegen, die speziell auf die individuellen Bedürfnisse zugeschnitten ist. Eine Auswahl ergänzend zum OECD-Katalog sei an dieser Stelle genannt:

Für Technikfolgenabschätzung in ethischer Dimension können Reflexionsmethoden wie das MEESTAR-Modell (Weber 2016, Ethical OS n. d.) oder Workshop-Formate, bspw. Ethical Explorer (Ethical Explorer n. d.), hilfreich sein. Workshops bieten die Möglichkeit, Stakeholder unter den relevanten Diversitätsgesichtspunkten miteinander ins Gespräch zu bringen und über die verschiedenen Projektphasen hinweg als multiperspektivischen Reflexionskreis, angepasst an das jeweilige Produktszenario, einzubinden. Andere Standards und Frameworks in Bezug auf beispielsweise Risikomanagement, Informationssicherheit (ISO/IEC 27001) und KI-Systeme (ISO/IEC 42001, ISO/IEC TR 24368, ISO/IEC 23053) können ergänzende Leitplanken für die Umfeldbeobachtung und Zielentwicklung stellen. Sollte die Anwendung von Prozessstandards nicht möglich sein, kann stattdessen auf *Playbooks* aufgesetzt werden, wie den *AlgoRules* (Algo.Rules n. d.) oder dem *etami open guidebook* (Etami open guidebook n. d.). Um Prozesse und Datenflüsse aufzuschlüsseln, können Programme, wie beispielsweise der *Data Process Modeler* (Bayern innovativ n. d.), eingesetzt werden. Die hier genannten Werkzeuge sind zum Teil umfangreich und erfordern mitunter ethische Expertise für einen effektiven Einsatz. Geeignete Werkzeuge sind deshalb kontextspezifisch (insbesondere nach Umfang und Aufwand entsprechend der Produkt- und Prozessgegebenheiten) auszuwählen oder anzupassen.

Ebenso gibt es für die wertorientierte Ausrichtung von Unternehmen eine Vielzahl an Kodizes und Leitbildern, welche als Orientierungspunkt für die Reflexion der unternehmenseigenen Haltung dienen können. Als Inspiration genannt seien hier zum Beispiel die *Digital Responsibility Goals* (Meier et al. 2022) und der CDR-Kodex (Corporate Digital Responsibility Initiative 2023). Bei der Formulierung und Reflexion des eigenen Leitbildes sind partizipative *bottom-up*-Prozesse essenziell, um bereits bestehende Unternehmenswerte zu integrieren und der eigenen Unternehmenskultur angemessen Rechnung zu tragen. Darauf aufbauend lässt sich die Operationalisierung und Messung der Vertrauenswürdigkeit weiter optimieren, bspw.: von der unternehmensinternen Datenethik-Leitlinie (*Directive*) über interdisziplinäre Dialogformate zu Data Governance und Datenethik (*Datenethik Hub, Enterprise Data Council*) bis hin zu den konkreten internen Beratungs- und Unterstützungsleistungen mit

Blick auf datengetriebene Projekte (Datenkatalog; PWC 2023; DATEV 2023b, S. 7).

Die erste Phase ist entscheidend für den weiteren Verlauf der Entwicklung und den Aufbau von Vertrauenswürdigkeit. Ohne eine geordnete, klare und nachvollziehbare Aufarbeitung der Rahmenbedingungen, sind Ziele nur schwer zu formulieren und eine vertrauenswürdige Entwicklung schwer zu kommunizieren. Im gleichen Zug ermöglicht diese tendenziell explorative Phase eine Konkretisierung der Produktleistungen (Hauer et al. 2023). In Bezug auf das Vertrauen wird erkundet, wer die potenziell vertrauensgebende Instanz ist. Außerdem wird festgestellt, mit welchen Maßnahmen die vertrauensnehmende Instanz die Vertrauenswürdigkeit der Produkte stärken kann.

Phase 2: Maßnahmenplanung und -durchführung – *Ethical Value Requirements*

Auf Basis der festgelegten Ziele und Bedingungen gilt es im nächsten Schritt konkrete Maßnahmen zu planen und durchzuführen. Funktionalitäten und Systemeigenschaften (*Ethical Value Requirements*, EVR *[IEEE SA. 2021, 18]*) werden unter Berücksichtigung der Werte aus Phase 1 bestimmt und tragen diese in die praktische Umsetzung. Bei bestehenden Produkten werden an dieser Stelle nötige systemische Anpassungen gefunden, um fortlaufend auf ein einheitliches und vertrauenswürdiges Produktportfolio einzuzahlen. Um die Umsetzung zu gewährleisten, sollten für diese Phase ebenfalls geeignete Dokumentationswerkzeuge und Kommunikationsmaßnahmen bestimmt werden. In Bezug auf wertebasierte Produktentwicklung können ergänzende *Ethical Checkboxes* in bereits etablierte Checklisten und Projektprozesse integriert werden. Aus dem Umfeld der *Data Science* sei hier beispielhaft das Werkzeug *deon* (deon n. d.) genannt, womit sogenannte *Ethical Checkpoints* (ECP) in datengetriebenen Projekten erstellt werden können. Es kann hilfreich sein, auch in dieser Phase Workshop-Formate zu nutzen, um die Suffizienz

der verabschiedeten Maßnahmen aus Sicht verschiedener Stakeholder beurteilen zu können.[11]

Für die Maßnahmenplanung und -durchführung greift DATEV bei Bedarf auf Dialogformate zu wertorientierter Data Governance, wie den *Enterprise Data Council* oder den Datenethik-Hub, zurück. In regelmäßigen Terminen und einzelnen Workshops werden die priorisierten Werte aus Phase 1 in konkrete Anforderungen und Maßnahmen übersetzt. Durch die Dokumentation in gemeinsamen Arbeitsumgebungen ist es möglich, diese zu reflektieren und regelmäßig zu überprüfen. Der Datenkatalog trägt dazu bei, dass sich Anfrageprozess und Datennutzung nachvollziehen lassen. Dies ist eine wichtige strukturelle Voraussetzung für *first party audits* oder Tests. In diesem Sinne findet ein Beitrag zu den Parts *Integrity* und *Ability* des ABI-Modells statt.

Phase 3: Maßnahmenbewertung und Iteration – Qualitätssicherung der *Ethical Value Requirements*

Zur Bewertung der erfolgten Maßnahmen, wird der Erfüllungsgrad von wertebasierten Zielen geprüft. Hierzu können Aktivitäten aus Phase 1 herangezogen werden, wie die Ergebnisse der Technikfolgenabschätzung. Des Weiteren können externe Instanzen für Audits oder Zertifikate hinzugezogen werden. Erreicht das geschaffene Produkt nachweislich und für die Stakeholder nachvollziehbar die gesteckten Ziele, dann stärkt dies die Vertrauenswürdigkeit des Vertrauensobjekts und in der Regel auch die der vertrauensnehmenden Instanz, insbesondere wenn die Technikgestaltung in Orientierung an den gesetzten Werten erfolgt ist (Integrität). Wurde ein Kriterienkatalog angelegt, kann dieser auf Erfüllung – intern und/oder extern – geprüft werden. Sollte die Erfüllungsrate geringer ausfallen als gewünscht, gibt er Orientierung für die Fokuspunkte der weiteren normativen Produktentwicklung. Bei ausreichender Erfüllungsrate kann der Kriterienkatalog selbst und der Grad der Erfüllung geprüft oder zertifiziert und als Vertrauensgrundlage an die vertrauensgebende Instanz kommuniziert werden.

11 Ein weiteres mögliches Ergebnis dieser Phase ist ein anwendungsbezogener Kriterienkatalog, welcher konkrete Anforderungen an das Produkt stellt. Dieser kann als Basis für externe Prüfungen genutzt werden, um die Qualität des Produkts von außen zu bestätigen. Um vage Metriken zu vermeiden, ist eine Finalisierung der Kriterien in Phase 2, statt schon in Phase 1, vorzuziehen.

Im Sinne des ABI-Modells kann durch Tests unter Berücksichtigung der Produktleistung die Ability der vertrauensnehmenden Instanz untersucht werden. Eine hohe Leistung lässt sich im Kontext des geprüften Produkts intersubjektiv nachvollziehbar extern kommunizieren. Gleichzeitig kann zusätzliche wertorientierte Weiterentwicklung die Integrität des Vertrauensnehmers fördern. Über Szenarienkataloge lassen sich gebündelte Kriteriensets prüfen und Aussagen über die Vertrauenswürdigkeit für spezifische Szenarien treffen.[12]

Die Directive Datenethik lässt sich im Kontext der Maßnahmenbewertung für den genannten Use Case als Möglichkeit heranziehen, wertorientierte Anforderungen an datengetriebene Produkte interdisziplinär zu überprüfen. Die o.g. Dialogformate kommen u.a. dann zum Einsatz. Hierbei referenziert der Prozess auf die Phase 1, insofern Potenziale mit Blick auf operationalisierte Werte identifiziert werden. Dieses Vorgehen bietet eine Basis, um eine perspektivische Auditierung strukturiert vorzubereiten.

Zusammenfassung und Fazit

Der hier vorgeschlagene Weg, Vertrauen in datengetriebene Projekte aufzubauen und zu festigen, verknüpft konzeptionelle Überlegungen mit einem praxistauglichen Lösungsansatz. Ausgehend von analytischen Vertrauensmodellen und Ansätzen aus der sozialwissenschaftlichen Organisationsforschung skizziert dieser Beitrag ein bedarfs- und anwendungsorientiertes Werkzeug, um Vertrauen zu messen und zu operationalisieren. Mit Blick auf den inflationären und zumeist vagen Gebrauch des Begriffs Vertrauen grenzt dieser Beitrag den Untersuchungsgegenstand zunächst auf das Konzept der Vertrauenswürdigkeit von konkreten Datennutzungsszenarien ein. Im anschließenden Anwendungsszenario dient ein wertorientierter fachlicher Datenkatalog zur Operationalisierung einer vertrauenswürdigen Datennutzungspraxis. Der nächste Schritt erörtert problemorientiert die Möglichkeiten von Vertrauensmessung. Abschließend zeigt der Beitrag, welche weiteren Werkzeuge sich zur Sicherstellung einer vertrauenswürdigen

12 Um die Vertrauenswürdigkeit des Produkts für verschiedene Nutzungsprofile differenziert zu betrachten, werden die jeweils relevanten Kriterien in Kriteriensets zusammengefasst. Beispielsweise kann zwischen B2B- und B2C-Nutzung unterschieden werden, da sich die beiden in Nutzungsgrundlage und -umständen unterscheiden.

Datennutzungspraxis anbieten und wie diese zu einer kontextspezifischen Erweiterung des Datenkatalogs führen können.

Zusammenfassend lässt sich festhalten, dass sich die Operationalisierung und Messung von Vertrauen nicht allein über Akzeptanzraten oder -geschwindigkeiten steuern lässt. Die Kombination von internen und externen Tests, Audits und Zertifizierungen mit kriterienbasierten Ansätzen bietet eine umfassendere Bewertung der Vertrauenswürdigkeit. Dabei bleibt der Aspekt »Kontext« entscheidend: So geht es im Rahmen dieses Beitrags nicht darum, ein basales Urvertrauen zu begründen oder eben genau unbegründet zu lassen. Sondern es geht darum, Vertrauen in seiner Dimension als Entscheidungsvertrauen zu verorten. Aus wirtschaftlicher Sicht betrachtet: Kunden entscheiden sich, unter Bezug auf ihre aktuelle Einsicht in den jeweiligen Zusammenhang, für oder gegen einen bestimmten Service, ein Produkt oder eine Dienstleistung. Persönliche Vertrauensbeziehungen auf der Mikroebene stärken das Vertrauen in bestimmte Unternehmen oder Institutionen auf der Mesoebene. Dies trägt wiederum dazu bei, das allgemeine Systemvertrauen auf der Makroebene aufrechtzuerhalten. Dieser Zusammenhang erklärt die oben ausgeführte hohe Bedeutung von Stakeholder-Einbezügen über unterschiedliche Formate. Für die IT-Branche hat dies zur Folge, dass insbesondere eine vertrauenswürdige Datennutzungspraxis im Rahmen der aktuellen Technikgestaltung essenziell für das Vertrauen in die Zukunft datengetriebener Geschäftsmodelle ist. Es ist die Voraussetzung dafür, Mehrwert aus Daten gemeinsam auf einem ethischen Fundament zu gestalten. Zukünftige Forschung und Entwicklungen sollten darauf abzielen, anwendungsorientierte und kontextspezifische Lösungen für die Vertrauensbewertung zu entwickeln und gesellschaftliche wie auch unternehmerische Belange miteinbeziehen. Mit Blick auf die Wertorientierung gilt es eine angemessene Balance zwischen innovativer Technikgestaltung und angemessener Regulierung zu finden.

Literatur

Ada Lovelace Institute. 2023. »Ada in Europe.« Accessed July 18, 2023. https://www.adalovelaceinstitute.org/our-work/europe/.

Al-Ruithe et al. 2019. »A systematic literature review of data governance and cloud data governance.« In: personal and ubiquitous computing, 2019, Vol.23 (5–6), p.839-859. https://doi.org/10.1007/s00779-017-1104-3.

Anzinger, Heribert M. 2021. Corporate Digital Responsibility. In: Nietsch, Michael (Hg.): Corporate Social Responsibility Compliance. München, S. 611–633.

Algo.Rules. N. d. »Regeln für die Gestaltung algorithmischer Systeme.« Accessed July 18, 2023. https://algorules.org/de/startseite.

Bayern Innovativ. N. d. »Data Process Modeler. Wie aus Transparenz Wettbewerbsvorteile werden.« Accessed July 18, 2023. https://www.bayern-innovativ.de/de/seite/data-process-modeler.

Benrath, Bastian. 2019. »Die wichtigste Ressource für Tech-Unternehmen. Vertrauen der Nutzer.« In Frankfurter Allgemeine Zeitung. Last modified September 19, 2019. https://www.faz.net/aktuell/wirtschaft/diginomics/vertrauen-der-nutzer-wichtigste-ressource-fuer-tech-unternehmen-16386248.html?GEPC=s5.

Bundesnetzagentur. 2018. »Verordnung (EG) Nr. 765/2008«. Accessed July 18, 2023. https://www.bundesnetzagentur.de/SharedDocs/Downloads/DE/Sachgebiete/Telekommunikation/Unternehmen_Institutionen/Technik/DMUEF/VO_EG_765_2008.html.

Capgemini Digital Transformation Institute survey. 2017. »The key to Loyalty.« Online unter: https://www.capgemini.com/wp-content/uploads/2017/11/dti_loyalty-deciphered_29nov17_final.pdf.

Corporate Digital Responsibility Initiative. 2023. »Kodex.« Accessed July 18, 2023. https://cdr-initiative.de/kodex.

Datenschutzkonferenz (DSK). 2019. »Akkreditierungsprozess für den Bereich ›Datenschutz‹ gemäß Art. 42, 43 DS-GVO.« https://www.datenschutzkonferenz-online.de/media/oh/20190315_oh_akk_c.pdf.

Datenschutzkonferenz (DSK). 2020. »Anforderungen zur Akkreditierung gemäß Art. 43 Abs. 3 DS-GVO i. V. m. DIN EN ISO/IEC 17065«. Last modified October 8, 2020. https://www.datenschutzkonferenz-online.de/media/ah/20201008_din17065_Ergaenzungen_deutsch_nach_opinion.pdf.

DATEV eG. 2021. DATEV Directive Datenethik. Online zukünftig unter: https://www.datev.de/web/de/ueber-datev/das-unternehmen/corporate-responsibility/.

DATEV eG. 2023a. DATEV-Verhaltenskodex – Code of Business Conduct. Online unter: https://www.datev.de/web/de/m/ueber-datev/das-unternehmen/compliance/verhaltenskodex/.

DATEV eG 2023b. Whitepaper »Datenschutz und Unternehmenssicherheit bei DATEV. Online unter: https://www.datev.de/web/de/m/ueber-datev/datenschutz/.

deon. N. d. »An ethics checklist for data scientists.« Accessed July 18, 2023. https://deon.drivendata.org/.

Deutscher Ethikrat (Hg.). 2023. »Mensch und Maschine – Herausforderungen durch Künstliche Intelligenz.« Stellungnahme. Accessed July 18, 2023. https://www.ethikrat.org/fileadmin/Publikationen/Stellungnahmen/deutsch/stellungnahme-mensch-und-maschine.pdf.

etami open guidebook. N. d. »The open guidebook on legal, trustworthy, and ethical Artificial Intelligence.« Accessed July 18, 2023. https://guidebook.etami.org/.

Ethical Explorer. N. d. »Ethical Explorer.« Accessed July, 2023. https://ethicalexplorer.org/.

Ethical OS. N. d. »Ethical OS Toolkit«. Accessed July 18, 2023. https://ethicalos.org/.

European Commission. 2021. »Proposal for a Regulation of the European Parliament and of the Council laying down perationa rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts.« Accessed July 18, 2023. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206.

Filipović, Alexander. 2023. »Corporate Digital Responsibility« (Preprint). In The Routledge Handbook of Responsibility, edited by Maximilian Kiener. New York, London.

Fraunhofer IAIS. 2023. »Künstliche Intelligenz sicher und vertrauenswürdig gestalten.« Accessed July 18, 2023. https://www.iais.fraunhofer.de/de/forschung/kuenstliche-intelligenz/ki-zertifizierung.html.

Gerlach, Robert. 2019. »The Digital Product Ethics Canvas.« Accessed July 18, 2023. https://www.threebility.com/post/the-digital-product-ethics-canvas.

Hallensleben, Sebastian, Carla Hustedt, Lajla Fetic, Torsten Fleischer, Paul Grünke, Thilo Hagendorff et al. 2020. »From Principles to Practice. An interdisciplinary framework to operationalize AI ethics.« Edited by VDE and Bertelsmann Stiftung. AI Ethics Impact Group. https://www.ai-ethics-impact.org/en.

Hauer, Marc, Lena Müller-Kress, Gertraud Leimüller, Katharina Zweig. 2023. »Using Assurance Cases to assure the fulfillment of non-functional requirements of AI-based systems – Lessons learned.« In IEEE International Conference on Software Testing, Verification and Validation Workshops (ICSTW). DOI: 10.1109/ICSTW58534.2023.00040.

IEEE SA. 2021. »IEEE Standard Model Process for Addressing Ethical Concerns during System Design« Accessed July 18, 2023. https://standards.ieee.org/ieee/7000/6781/.

ISO/IEC 17021–1:2015 Konformitätsbewertung – Anforderungen an Stellen, die Managementsysteme auditieren und zertifizieren.

ISO/IEC 17024:2021 Konformitätsbewertung – Allgemeine Anforderungen an Stellen, die Personen zertifizieren.

Janssen, Marijn et al. 2020. »Data governance: Organizing data for trustworthy Artificial Intelligence.« In: Government information quarterly. Vol.37 (3), p.101493, Article 101493. https://doi.org/10.1016/j.giq.2020.101493.

Kaminski, Andreas. 2017. »Hat Vertrauen Gründe oder ist Vertrauen ein Grund? – Eine (dialektische) Tugendtheorie von Vertrauen und Vertrauenswürdigkeit.« In Praxis und ›zweite Natur‹, edited by Jens Kertscher and Jan Müller, S. 167–188. Brill | mentis. https://doi.org/10.30965/9783957438249_017.

Koska, Christopher. 2023. Ethik der Algorithmen. Auf der Suche nach Zahlen und Werten. (1. Auflage 2023.) Berlin: Springer Berlin; J.B. Metzler (Techno:Phil – Aktuelle Herausforderungen der Technikphilosophie, 6). https://doi.org/10.1007/978-3-662-66795-8.

Koska, Christopher, Julian Prugger, Sophie Jörg, Michael Reder. 2024. Die Verlagerung von Vertrauen vom Mensch zur Maschine. Eine Erweiterung des zwischenmenschlichen Vertrauensparadigmas im Kontext Künstlicher Intelligenz. Zeitschrift für Praktische Philosophie (ZfPP).

Leitfaden zur Auditierung von Managementsystemen. (ISO 19011: 2018); Deutsche und Englische Fassung EN ISO 19011:2018. Note 3.1.

Luhmann, Niklas. 1968. Vertrauen. Ein Mechanismus der Reduktion sozialer Komplexität. Stuttgart: Enke.

McAllister, D. J. 1995. »Affect- and cognition-based trust as foundations for interpersonal cooperation in organizations.« In Academy of Management Journal 38 (1), S. 24–59. DOI: 10.2307/256727.

Meier, Jutta, Hermsen, Kai, Bauer, Jochen und Bjoern Eskofier. 2022. Digital Responsibility Goals – A Framework for a Human-Centered Sustainable Digital Economy with a Focus on Trusted Digital Solutions. DOI: 10.3233/SHTI220377.

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

OECD.AI Policy Observatory. N. d. »Catalogue of Tools & Metrics for Trustworthy AI.« Accessed July 18, 2023. https://oecd.ai/en/catalogue/tools?terms=&page=1.

Online Ethics Canvas. N. d. »Online Ethics Canvas.« Accessed July 18, 2023. https://www.ethicscanvas.org/.

Open Data Institute. N. d. »What is the Data Ethics Canvas?«. Accessed July 18, 2023. https://theodi.org/article/the-data-ethics-canvas-2021/.

PWC (Hg.). 2023. Digitale Ethik & Verantwortung. Wie eine langfristige wirkungsvolle Verankerung in ihrem Unternehmen gelingt.

Schoorman, F. David, Roger C. Mayer, James H. Davis. 2007. »An Integrative Model of Organizational Trust: Past, Present, and Future.« In The Academy of Management Review 32 (2), S. 344–354. DOI: 10.5465/amr.2007.24348410.

Seidl, Martina. 2022. »Corporate Digital Responsibility: Stimulating Human Centric Innovation and Building Trust in the Digital World.« In: Jacob, Kai et al. (Hg.): Liquid Legal – Humanization and the Law. Cham, S. 55–81.

Slaper, Timothy F., Tanya J. Hall. 2011. »The Triple Bottom Line: What Is It and How Does It Work?« In Indiana Business Review 86 (1), S. 4–8. https://www.ibrc.indiana.edu/ibr/2011/spring/pdfs/article2.pdf.

TÜV-Verband. 2021. »Studie Künstliche Intelligenz.« Online unter: https://www.tuev-verband.de/?tx_epxelo_file%5bid%5d=856780&cHash=2a8f26ddede3cbe286f5d30efd9759dd.

van de Poel, Ibo. 2013. »Werthaltigkeit der Technik.« In Handbuch Technikethik, edited by Armin Grunwald und Melanie Simonidis-Puschmann. Stuttgart: J.B. Metzler, S. 133–137.

VDE. 2022. »VCIO based description of systems for AI trustworthiness characterisation. VDE SPEC 90012 V1.0 (en).« Accessed July 18, 2023. https://www.vde.com/resource/blob/2242194/a24b13db01773747e6b7bba4ce20ea60/vcio-based-description-of-systems-for-ai-trustworthiness-characterisationvde-spec-90012-v1-0--en--data.pdf.

Verma, Sahil, Julia Rubin. 2018. »Fairness definitions explained.« In Proceedings of the International Workshop on Software Fairness – FairWare '18. the International Workshop, edited by Yuriy Brun, Brittany Johnson und Alexandra Meliou, S. 1–7. Gothenburg, Sweden, 29.05.2018 – 29.05.2018. New York, New York, USA: ACM Press.

Wagenblass, Sabine. 2018. »Vertrauen.« In Handbuch Soziale Arbeit. Grundlagen der Sozialarbeit und Sozialpädagogik (6., überarbeitete Auflage)«, edi-

ted by Hans-Uwe Otto, Hans Thiersch, Rainer Treptow und Holger Ziegler. München: Ernst Reinhardt Verlag.

Weber, Karsten. 2016. »MEESTAR² – Ein erweitertes Modell zur ethischen Evaluierung soziotechnischer Arrangements.« Conference Paper: Zweite transdisziplinäre Konferenz zum Thema »Technische Unterstützungssysteme, die die Menschen wirklich wollen« 2016. https://www.researchgate.net/publication/311699459_MEESTAR_-_Ein_erweitertes_Modell_zur_ethischen_Evaluierung_soziotechnischer_Arrangements.

ZVKI. 2022. ZVKI-Online-Befragung: »Wissen, Nachvollziehbarkeit und bewertbare Erfahrungen – Zutaten für vertrauenswürdige Künstliche Intelligenz (KI).« Online unter: https://www.zvki.de/zvki-exklusiv/fachinformationen/online-befragung

Humaner als der Mensch? Zur sozialen Imagination autonomer Waffentechnik

Nicole Kunkel

1. Einleitung

Die Terminator-Reihe gehört zu den einflussreichsten und populärsten Science-Fiction-Filmen überhaupt. In ihrem ersten Teil wird ein Terminator, also eine anthropomorph gestaltete Maschine, aus einer dystopischen und von Maschinen regierten Zukunft in die Gegenwart der 1980er geschickt, um die junge Frau Sarah Connor zu töten. Sarah wird im Laufe ihres Lebens die Mutter von John Connor werden, dem Anführer des Widerstands gegen die Roboterübermacht in der Zukunft. Um eben diese Mutterschaft zu verhindern, ist es die Aufgabe des Terminators Sarah zu töten. Die Maschine setzt alles daran, kaltblütig ihren Auftrag auszuführen – sie achtet weder auf andere Opfer, noch kann sie durch die einfache Feuerkraft anderer Waffen gestoppt werden. Zwar gelingt Sarah die Flucht – so will es das eherne Gesetz des Actionfilms – doch ist die Maschine dem Menschen an Kaltblütigkeit und aufgrund seiner Unverwundbarkeit massiv überlegen: Weil der Android keinen Schmerz fühlt, sich nicht fürchtet und kein Erbarmen zeigt, ist er die ideale Killermaschine.

Umso faszinierender ist die Fortsetzung dieser Geschichte im zweiten Teil der Terminator-Reihe. Wieder wird ein Terminator aus der Zukunft geschickt – diesmal um den jungen John Connor zu töten. Dieser jedoch schickt selbst einen weiteren, umprogrammierten Terminator in die Vergangenheit, um sein eigenes Leben vor dem »bösen« Terminator zu schützen. Interessant ist hier das Verhältnis, das der umprogrammierte Terminator und John Connor im Laufe des Filmes zueinander aufbauen: Während die »gute« Maschine den Jungen gegen den »bösen« Terminator verteidigt, lernt sie nebenbei nicht nur flotte Sprüche von John, sondern auch allerlei Lektionen in Sachen Mitmenschlichkeit. So hinterlässt dieser Android keine Spur von Verwüstung, sondern

schießt beispielsweise gezielt daneben, sodass niemand zu Schaden kommt. Auch in dieser Fortsetzung siegt das Gute – der »böse« Terminator wird bezwungen und der Film endet dramatisch, indem der »gute« Terminator sich selbst opfert, da er sonst zum Prototyp für eben jene dystopische Technik der Zukunft genutzt werden könnte. Dieser Film folgt der Logik: Weil der Android mitmenschlicher ist als der Mensch, kann er sein »Leben« für die Menschen geben: Er ist der ideale, selbstlose Beschützer.

Obwohl beide Filme dieselbe Technik in Szene setzen, könnte die Moral der Geschichte unterschiedlicher kaum sein: Auf der einen Seite die kaltblütig mordende Killermaschine, die rücksichtlos tötet und damit der Imperativ, niemals solche Maschinen zu bauen; auf der anderen Seite der selbstlose Roboter, der Mitmenschlichkeit lernt und schließlich sich selbst opfert. Unterstrichen wird dies mit dem Epilog zum zweiten Teil aus dem Munde Sarah Connors:

> »The luxury of hope was given to me by the Terminator. Because if a machine can learn the value of human life… Maybe we can too.« (Terminator 2 1991)

Dieser Artikel will dem Gehalt dieser Worte Sarahs und der damit verbundenen Bilder in theologisch-ethischer Perspektive nachspüren: Es scheint mir alles andere als selbstverständlich, dass der Wert menschlichen Lebens von Technik, die gerade zum Töten konstruiert wurde, gelernt werden kann – zumal diese Technik den Sinn ihrer Unternehmung nicht versteht. Dies ist insbesondere da problematisch, wo solche fiktionalen Vorstellungen den gesellschaftlichen (und politischen) Umgang mit Technik, etwa letalen autoregulativen Waffensystemen, leiten und handlungswirksam werden, indem sie die politische Regulation von bestimmten Techniken und Verfahren entweder antreiben oder aber verhindern. Mein Hauptanliegen ist es hier aus theologischer Perspektive zu zeigen, dass die Vorstellung, gerade den Wert menschlichen Lebens von rechnender Technik zu lernen, fehl geht. Die zentrale These besteht darin, dass die Maschine nicht wissen kann, was ein Mensch ist, was es bedeutet zu leben und dass entsprechend die Idee, solch zentrale Werte von einer Maschine zu lernen, eine absurde Fantasie ist, die zugleich das Menschenbild »technisiert«. Dies jedoch bedeutet nicht, dass eine solche Fantasie nicht auch in politischen Handlungsoptionen relevant werden kann. Ziel ist es also zunächst zu zeigen, dass Bilder und Narrative den Umgang mit Technik maßgeblich beeinflussen und bestimmen, sodann deutlich zu machen, dass die Vorstellung von der mitmenschlichen Technik grundsätzlich fehl geht, und schließlich diesen technischen Imaginationen auf Grundlage

der theologischen Idee der Ebenbildlichkeit Gottes ein Korrektiv entgegenzustellen.

Um dieses Thema genauer zu untersuchen, werde ich zunächst in die Imaginationstheorie von Charles Taylor einführen und anhand dieser zeigen, warum die Bilder und Geschichten, die wir erzählen, auch unsere kollektiven Vorstellungen prägen und somit handlungsleitend werden können (2). In einem zweiten Schritt analysiere ich bestehende Technikimaginationen und befrage sie auf ihren Gehalt (3), um diesen schließlich mit der Rede von der Ebenbildlichkeit Gottes eine Alternative der menschlichen Selbsterschließung entgegenzustellen (4).

2. Soziale Imagination

Wie also tragen Filme wie der Terminator mit ihren Bildern und Narrativen zu einem gesellschaftlichen und politischen Umgang mit Technik, etwa dem Umgang mit autoregulativen Waffensystemen, bei? Ob, und wenn ja, auf welche Weise fiktionale Vorstellungen tatsächlich gesellschaftliche und politische Strukturen prägen, und somit Filme wie die Terminator-Reihe Handlungsmacht entfalten können, lässt sich mit der Theorie des Philosophen Charles Taylor in seinem Buch *Modern Social Imaginaries* veranschaulichen. Taylors maßgebliche Frage ist hier, wie die moralische Ordnung innerhalb westlicher Gesellschaften ihre Wirkung entfaltet und er beantwortet diese Frage mit seiner Theorie zum sozial Imaginären. Unter moralischer Ordnung versteht Taylor den selbstverständlichen moralischen Verständnishintergrund, vor dem ein Mitglied einer bestimmten sozialen Gruppe sich selbst erschließt. Dieser Raum und Rahmen ist bereits moralisch vorgeprägt, wie eine Art Landkarte, auf der ein Individuum moralische Orientierung sucht (Taylor 1992, 29f.): Einen Standpunkt außerhalb dieser Landkarte einzunehmen ist dem Individuum unmöglich, weil es sich dann nicht mehr verständlich machen könnte. Taylor beschreibt diese Zusammenhänge anschaulich in seinem Buch *Sources of the Self* mit folgenden Worten: »[A] person without a framework altogether would be outside our space of interlocution; he wouldn't have a stand in the space where the rest of us are. We would see this as pathological.« (Taylor 1992, 31). Es ist der Einzelnen also unmöglich aus dem moralischen Rahmen oder Horizont herauszutreten, ohne zugleich aus den selbstverständlichen sozialen Zusammenhängen herauszufallen. Taylors These lautet nun weiter, dass eben diese moralische Ordnung sich weniger in expliziten Imperativen

und Theorien ausdrückt, sondern vielmehr implizit in den Praktiken, Narrativen und Selbstverständlichkeiten verbirgt, die tagtäglich und unsichtbar gesellschaftliches Handeln bestimmen. Er schreibt: »The social imaginary is not a set of ideas; rather, it is what enables, through making sense of, the practices of a society« (Taylor 2004, 2). Und er führt weiter aus:

> »By social imaginary, I mean something much broader and deeper than the intellectual schemes people may entertain when they think about reality in a disengaged mode. I am thinking, rather, of the ways people imagine their social existence, how they fit together with others, how things go on between them and their fellows, the expectations that are normally met, and the deeper normative notions and images that underlie these expectations.« (Taylor 2004, 23)

Worum es Taylor hier geht, sind also nicht die ausgesprochenen moralischen Theorien und Ideen, sondern die mitschwingenden Vorstellungen, die sich in Bildern, Geschichten und Legenden niederschlagen und die zwischenmenschlich geteilt werden, um so den selbstverständlichen Legitimationsgrund für eine kollektive Praxis bereitzustellen (Taylor 2004, 23). In diesen Bildern verdichten sich Ansprüche an die anderen, ein geteiltes Verständnis dessen, was von diesen anderen in Alltagssituationen erwartet werden kann und was so überhaupt erst ermöglicht, kollektive soziale Praktiken zu entwickeln. Die Komplexität dessen wird deutlich, wenn bewusst gemacht wird, dass in Imaginationen auch immer eine Vorstellung davon enthalten ist, auf welcher Art und Weise eine Gesellschaft zusammengehört, was normativ gilt, wie Abläufe normalerweise geregelt sind und wie sie es sein sollten (Taylor 2004, 24). Taylor charakterisiert beispielsweise die Art, wie wir gehen, gestikulieren und uns bewegen als einen Ausdruck dafür, wie wir uns zueinander positionieren. Er beschreibt etwa die betonte Langsamkeit, mit der ein Polizist aus dem Streifenwagen aussteigt und sich einer Autofahrerin nähert, die er wegen Geschwindigkeitsüberschreitung angehalten hat (Taylor 1992, 15). Diese Szene wird nicht verständlich ohne das Wissen darüber, wie Menschen sich in Alltagssituationen begegnen, oder eine Vorstellung davon, mit welcher Handlungsmacht die Polizei ausgestattet ist. Ebenso sehr ist ein Wissen davon nötig, dass Geschwindigkeitsüberschreitungen verboten sind, und dass von einer Autofahrerin erwartet wird, sich an die Straßenverkehrsordnung zu halten. Nicht nur das: Polizist*innen sind selbstverständlich zu bestimmten Handlungen autorisiert, zu denen die anderen Bürger*innen nicht legitimiert

sind. Deswegen tragen sie bestimmte Kleidung und fahren bestimmte Autos. Aufgrund dieser Autorisierung sind sie überhaupt erst in der Lage eine andere Person wegen einer Geschwindigkeitsüberschreitung zur Rechenschaft zu ziehen – und genau deswegen ist es ihnen auch erlaubt, Gewalt anzuwenden. Dieses System funktioniert jedoch nur, weil alle davon wissen und über einen geteilten und selbstverständlichen Wissenshintergrund verfügen. In dieser einen Szene verdichtet sich also spezifisches Hintergrundwissen, angefangen von der Vorstellung davon, welche Handlungsmacht Polizist*innen zusteht, der Norm einer Geschwindigkeitsbeschränkung, bis hin zu einer abstrakten Vorstellung davon, wie wir als Gesellschaft zusammengehören und welche Normen dieser Zusammengehörigkeit zu Grunde liegen – oder zu Grunde liegen sollten.

Zentral ist dabei, dass die Beziehung zwischen Hintergrundvorstellungen und Praktiken keineswegs einseitig ist. Vielmehr bringen neue Praktiken neue Imaginationen hervor, während gleichsam neue Imaginationen neue Praktiken prägen, sodass sich die soziale Welt in einem stetigen Wandel befindet (Taylor 2004, 24). In einem solchen Prozess ermöglichen neue Praktiken wiederum neue Perspektiven, die dann die neue Praktik mit Sinn füllen. Das heißt, dass unsere moralischen Vorstellungen dem Status quo nicht unbedingt entsprechen müssen: Sie können diesen überschreiten, neue Visionen entwickeln, die dann wiederum neue Praktiken hervorrufen und so Stück für Stück eine Veränderung in der sozialen Imagination bewirken (Taylor 2004, 28). Auf diese Weise wird eine neue Perspektive handlungswirksam und, setzt sie sich durch, wird sie irgendwann zu einem selbstverständlichen Hintergrund, der kaum erwähnenswert erscheint (Taylor 2004, 29). In Worten des Theologen Florian Höhne:

> »Das sozial Imaginäre lässt […] nicht nur (neue) Praktiken initiieren. Neue Praktiken wirken – wie Taylor betont – auch in das sozial Imaginäre zurück. Neue Dinge, etwa neue Technologien verändern Praktiken, was wiederum neue Deutungen, Theorien und Vorstellungen schafft, die über die Zeit in das Selbstverständliche des sozial Imaginären einsickern können.« (Höhne 2022, 117)

Eine philosophische oder soziale Theorie ist dann die Abstraktion geteilter Praktiken einer bestimmten sozialen Gruppe (Taylor 2004, 29f). Für westliche Gesellschaften etwa weist Taylor auf die prägende Rolle von Öffentlichkeit,

oder aber die Herrschaft des Volkes hin, die jeweils mit bestimmten sozialen Imaginationen und Praktiken einher gehen.

Im Blick auf die hier angesprochenen Narrative und Bilder von Technik heißt das, dass der Terminator, auch wenn der Inhalt des Filmes rein fiktional ist, eine bereits vorgängige Intuition aufnimmt – sonst würde sich uns die Handlung des Filmes nicht erschließen und die Filme wären Ladenhüter geblieben. Zugleich aber wirkt die Imagination normativ auf unsere Vorstellungen für den Umgang mit hochentwickelter (Waffen-)Technik. Es ist die Vorstellung der kaltblütigen Maschine, die erbarmungslos tötet, die hier einerseits als soziale Selbsterschließung des Umgangs mit Technik gesehen werden kann. Zugleich verbirgt sich dahinter der Imperativ, niemals solche Technik herzustellen, wenn die Zukunft nicht entsprechend dystopisch aussehen soll. Dass diese sehr pessimistische Art der Darstellung sich mit der Maschine, die selbstloser ist als der Mensch, paart, macht das Narrativ im Film nur umso schillernder: Der Terminator opfert sich selbst, um die Menschheit zu schützen, denn sein Dasein gefährdet deren Zukunft.

Greifbar werden beide Narrative gegenwärtig in der Debatte um autoregulative Waffensysteme, wenn auf der einen Seite die möglichen Vorteile dieser Maschinen gepriesen werden, die letztlich menschlicher als der Mensch funktionieren sollen (Arkin 2009; Müller 2016). Autor*innen, die so argumentieren, knüpfen damit an das Narrativ der selbstlosen Maschine an, die ihr »Leben« für den Menschen geben würde. Auf der anderen Seite verbinden Gegner*innen der Technik ihre Darstellung nicht selten mit dem Sprachbild des *Killerroboters*, was keineswegs nur als eine neutrale Beschreibung der Technik gesehen werden muss, wie etwa die *Campaign to stop killer robots* bereits in ihrem Namen deutlich macht (Campaign to stop killer robots n.d.). *Killerroboter* sind der Inbegriff der kaltblütig tötenden Maschine, die rücksichtslos tötet, was ihr in den Weg kommt. In diesen entgegengesetzten Narrativen spiegelt sich gleichsam das höchst ambivalente Verhältnis zu Automatisierung und Digitalisierung in anderen Zusammenhängen. Dass sowohl das Narrativ der selbstlosen Maschine wie das des *Killerroboters* durch die evozierten Bilder und Vorstellungen gesellschaftliche Vorstellungen aufnehmen und transformieren, sollte deutlich geworden sein. Entsprechend ist die Idee, auf die Bedrohung durch diese Systeme mit dem Terminus des *Killerroboters* hinzuweisen, weder besonders originell noch erklärungsbedürftig, ebenso wenig wie die Hoffnung, mit neuer Technik das Leben der Menschen zu verbessern, bzw. zu schützen. Was sich daran aber zugleich zeigt, ist ein Komplex an anthropologischen Fragestellungen, die viel stärker auf das Selbstverständnis des Menschen zielen.

3. Paradoxe Technikimaginationen

Kommen wir noch einmal zur eingangs erwähnten Erklärung Sarah Connors zurück, die festhält: »[I]f a machine can learn the value of human life.... Maybe we can too.« (Terminator 2 1991). Diese Bemerkung ist aus verschiedenen Gründen aufschlussreich. Zunächst einmal ist sie höchst paradox, denn dass ein Terminator, dessen namengebende Aufgabe es ist, das Leben von Menschen zu beenden (lat. *terminare* – beenden), gerade den Wert dieses Lebens schätzen soll, ist ein Widerspruch in sich selbst. Zweitens ist es keineswegs selbstverständlich anzunehmen, dass eine Maschine so etwas wie den Wert des Lebens zu schätzen weiß. Wo das angenommen wird, wird entweder anhand menschlicher Vorbilder auf ein nicht vorhandenes Innenleben der Maschine zurückgeschlossen, ein Vorgang den Frederike van Oorschot als imitative Imagination ausgewiesen hat (van Oorschot 2023) – oder aber eine Simulationsperfektion vorausgesetzt, die alles andere als trivial ist.[1] Drittens – und darauf kommt es hier an – geht Sarah davon aus, dass die Maschine dem Menschen ein Vorbild in Bezug auf ihre ethische Kompetenz werden kann, denn im Gegensatz zu den Menschen, weiß die Maschine ja anscheinend nicht nur besser, was der Wert des Lebens eigentlich ist, sondern kann auch noch besser als der Mensch danach handeln. Entsprechend ist die dramatische »Selbsttötung« des Terminators am Ende des Filmes folgerichtig.

Dass diese Vorstellungswelten nicht nur unter Science-Fiction-Autor*innen verbreitet sind, zeigt ein Blick in die Forschungslandschaft zu autoregulativen Waffensystemen, wo beispielsweise der Robotiker Ronald C. Arkin davon ausgeht, dass dank der fortschreitenden Technik Kriege humaner ausgetragen werden könnten. Er schreibt hierzu:

> »Roboter sind bereits jetzt schneller, stärker und in einigen Fällen (z.B. *Deep Blue, Watson*) intelligenter als Menschen. Warum fällt es uns dennoch so

1 Bemerkt sei an dieser Stelle, dass ja auch der Mensch, bzw., die kognitiven Prozesse des Menschen, wie Intelligenz, Lernen, Autonomie, Urteilen hier simuliert werden sollen. Trotz der anthropomorphisierenden Sprechweise aber fallen die kognitiven Prozesse des Menschen mit den Rechenoperationen der Maschine nicht in eins. Auch wenn der Mensch diese ihm eigenen Prozesse in der Maschine wiederzuentdecken meint, handelt es sich dabei um bloße Simulationen. An Stellen aber, wo davon ausgegangen wird, dass eine solche Simulation in einer (nicht) näher spezifizierten Zukunft tatsächlich dem Menschen zum Verwechseln ähnlich ist, bleiben viele technische Fragen offen.

> schwer uns vorzustellen, dass sie irgendwann fähig sein könnten, uns auf dem Schlachtfeld menschlicher zu behandeln als wir Menschen selbst – obwohl sich menschliche Kampfkräfte in diesen Situationen immer wieder furchtbare Grausamkeiten zuschulden kommen lassen?« (Arkin 2014, 5).

Der Philosoph Vincent C. Müller kommt zu einem ähnlichen Schluss, wenn er autoregulative Waffensysteme langfristig als eher gute Neuigkeiten beschreibt und für ihn klar zu sein scheint, dass solche Systeme die menschlichen Kosten im Krieg verringern würden (Müller 2016, 78). In diesen wissenschaftlichen Ausführungen scheint das Narrativ der Maschine, die menschlicher ist als der Mensch zu greifen. Ist dem so, dann liegt es nahe, wie Arkin dies tut, davon auszugehen, dass es einen moralischen Imperativ zur Erforschung und Nutzung solcher Technologien gibt (Arkin 2014, 5).

Wie bereits beschrieben, ist jedoch genau diese Annahme äußerst voraussetzungsreich und an dieser Stelle sogar explizit erweitert um die Unfähigkeit des Menschen selbst moralisch integer zu handeln, allem voran auf dem Schlachtfeld. Um dies zu konstruieren, geht Arkin in seinen Beiträgen umfassend auf die moralischen Unzulänglichkeiten des Menschen ein, die vor allem auf dem Schlachtfeld zu Tage treten. Ob es die Lust an Rache, die Notwenigkeit zum Selbstschutz, das Töten aus Angst, Müdigkeit oder schiere Überforderung ist: die Maschine scheint die bessere, weil rationalere Wahl zu sein (Arkin 2014, 5f; ders. 2009, 29–35). Deswegen, so die Argumentation, kommt es zu weniger Opfern. Im Vergleich mit der Maschine, schneidet der Mensch hier also schlechter ab.

Der Vergleichspunkt, an dem sich Arkin hier orientiert, stammt aus Normen und Pflichten der Mitmenschlichkeit. Diese haben ihre juristische Ausprägung im internationalen Völkerrecht gefunden, etwa der Genfer Konvention, die zwischen Kombattant*innen und Nicht-Kombattant*innen unterscheiden, und den Schutz von Nichtkombattant*innen, wie verwundeten Soldat*innen, oder Zivilist*innen einfordern. Wer nach diesen Spielregeln spielt, so die Idee, verhält sich menschlich. Schieben wir das Problem, dass solche Waffen gar nicht zwischen Kombattant*innen und Nicht-kombattant*innen unterscheiden können für einen Moment zur Seite (Geiß und Lahmann 2017) und wenden den Blick auf die Argumentationsfigur und das dahinterliegende Narrativ, so zeigt sich ein Phänomen, das Günther Anders als prometheische Scham bezeichnet. Gemeint ist die »Scham vor der ›beschämend‹ hohen Qualität der selbstgemachten Dinge« (Anders 1956, 23). Anders meint damit die Erkenntnis, dass das eigene Leben, der eigene Leib, ja, das

eigene Selbst im Gegensatz zur perfekten Maschine Makel aufweist. Da die Maschine angesichts des fehlerhaften Menschen als perfekt erscheint, gipfelt die prometheische Scham in der Selbstaufgabe des Menschen, nach dem Leitspruch: »Da wir schlechter rechnen als unser Apparat, sind wir unzurechnungsfähig; ›rechnen‹ wir also nicht« (Anders 1956, 61). Weil der Mensch hier an seine eigenen Grenzen zu stoßen scheint, gibt er »unzurechnungsfähig« die Kontrolle vertrauensvoll an die Maschine ab.

Das Problem, das einer solchen Erzählung zu Grunde liegt, weist Florian Höhne in seinem Aufsatz als Reduktion des Menschenbildes auf den risikoinformierten Entscheider aus: Der Mensch vergleicht sich mit der Maschine, allerdings nimmt er zum Maßstab die Fähigkeit Entscheidungen zu treffen, die eine gewisse Risikokalkulation beinhalten. Da die Maschine aber schneller und präziser rechnet, schneidet der Mensch hier schlechter ab, ohne dass jedoch andere Kompetenzbereiche in den Blick genommen werden (Höhne 2022). Mein Interesse hier gilt aber nicht so sehr der Auswirkung, sondern vielmehr dem Grund für eine solche reduktive Annahme. Und dieser zeigt sich hier in der Scham vor der Technik: Durch die zugrundeliegenden Anthropomorphisierungsprozesse (Kunkel 2023), vermittelt durch Filme wie den Terminator, identifiziert sich der Mensch mit der Technik – und sieht sich dann im Vergleich mit der Maschine an der gestellten Aufgabe scheitern. Das Problem aber ist: Hier werden Birnen mit Äpfeln verglichen, denn sachlich gibt es keinen Grund zur Gleichsetzung von Mensch und Maschine, besonders in ethischen Fragestellungen. Das liegt vor allem daran, dass für moralische Handlungen eine Vorstellung vom Handlungskontext, ein Welt- und Selbstbild nötig ist. In philosophischen Termini: ein autonomes Selbst. Eine Maschine, auch, wenn sie mit Verfahren sogenannter künstlicher Intelligenz arbeitet, hat ein solches Selbst jedoch nicht. Sie hat auch keine Vorstellung von der Welt und allem, was sich darin befindet. In Worten des Philosophen Brian Cantwell Smith: sie rechnet, urteilt aber nicht (Smith 2019). Um moralisch zu handeln, braucht es aber ein Selbst- und Weltbild, um zu verstehen, was es bedeutet sich in dieser Welt handelnd zu bewegen und aufgrund dessen (moralische) Urteile zu fällen (Kunkel 2021). Maschinen, die ihren Weltzugang jedoch berechnen, können dies nicht, denn sie haben kein Verständnis davon, was sie berechnen. Kurzum: Sie wissen nicht, womit sie es zu tun haben. Nichtsdestotrotz ist das Motiv des mitmenschlichen Terminators handlungswirksam, weil es diese Identifikationsprozesse ermöglicht, informiert und aufrechterhält.

Etwas anders gelagert ist die Problematik im entgegengesetzten Narrativ von der kaltblütig mordenden Maschine, wie sie etwa im Terminus des *Killerroboters* Ausdruck findet. Hier ist das Narrativ an der dystopischen Zukunft der Terminator-Reihe angelehnt und bedient die Angst vor der Machtübernahme durch Technik, weil sie gerade nicht in der Lage ist, ihre implementierten Rechenoperationen zu verstehen. Eine solche Technik kann kaltblütig töten, eben weil sie nicht weiß, was Leben ist und was Sterben bedeutet. Letztlich jedoch treten auch in dieser Imagination dieselben schamhaften Mechanismen hervor: insofern die Gegner*innen autoregulativer Waffentechnik darauf bestehen, dass solche Waffen niemals hergestellt werden dürfen und dabei auf die potentielle Vernichtung der Menschheit durch solche Systeme verweisen (Future of Live, n.d.) ist zugleich die Idee der übermenschlichen und nicht mehr zu stoppenden Maschine beschworen.[2] Das schamhafte Moment und damit der Impuls die Tötungsentscheidung an die Maschine abzutreten tritt hier gegenüber einem ungeheuerlichen, angsteinflößenden zurück. Dieser jedoch wirkt, als wäre der Mensch dem handlungsunfähig ausgesetzt. Etwas überspitzt formuliert: Wenn solche Systeme erst einmal gebaut werden dürfen, dann übernehmen Killerroboter die Macht. Dass auch hinter der Entwicklung und dem Einsatz von Waffensystemen jeder Art Menschen stehen, die mit bestimmten Motiven diese Waffen herstellen und vertreiben, kann dabei schnell aus dem Blick geraten. Interessant scheint mir aber, dass die »Scham vor der ›beschämend‹ hohen Qualität der selbstgemachten Dinge« (Anders 1956, 23) auch in diesem Narrativ erhalten bleibt – auch wenn sie in einem diametral entgegengesetzten Imperativ mündet, nämlich in der Forderung, diese Waffen gar nicht erst herzustellen, gerade weil sie, einmal in Gang gesetzt, nicht mehr zu stoppen sein könnten.

4. Der Mensch im Spiegel Gottes

Das Problem, das ich in den beschriebenen Zusammenhängen sehe, ist letztlich ein anthropologisches: Indem der Mensch sich mit der Maschine iden-

2 Gleichzeitig ist der Gedanke nicht von der Hand zu weisen, dass solche Waffen sich verselbstständigen, denn genau da scheint das Ziel einer autoregulativen Waffe zu sein. Einmal losgelassen jedoch könnte sie einen einzelnen Fehler innerhalb kürzester Zeit in rasanter Geschwindigkeit wiederholen – mit potentiell fatalen Folgen (Scharre 2016).

tifiziert und vergleicht, verändert sich gleichsam seine Selbstwahrnehmung, die jedoch letztlich einem Trugschluss aufsitzt. Am Beispiel des Terminators heißt das, dass der Versuch Mitmenschlichkeit mit Hilfe von Robotik zu erlernen zum Scheitern verurteilt sein muss, denn die Maschine ist ja gerade nicht menschlich, kann also auch nicht mitmenschlich »handeln« und die Imagination, dass dies so sei, ist eine Fehlinterpretation der Technik.

Wo dies aber dennoch geschieht, maschinisiert der Mensch nicht nur sich selbst (Fuchs 2020; Koch 2016), sondern übersieht zugleich, dass Technik nicht unabhängig vom Menschen existiert – genauso wenig wie Sprache (Coeckelbergh 2017). Auf den Punkt gebracht heißt das: Der Terminator hat kein Eigenleben, weil er immer schon mit Daten operiert, die in der Lebenswelt des Menschen fußen. Nicht nur das, er ist auch mit einer spezifischen Intention ausgestattet – in diesem Fall zu töten – und auf eine gewisse Art und Weise gestaltet, die diesem Ziel entspricht. Wo einer solchen Entität ein Eigenleben unterstellt wird, das dann wieder lehrreich für den Menschen sein kann, gerät zum einen die Intention der Designer*innen, Programmierer*innen oder Betreiber*innen aus dem Blick – und damit zugleich die Frage nach den dahinterliegenden Machtstrukturen, denn ein bestimmtes Gerät wird ja stets zu einem gewissen Zweck hergestellt. Zum anderen – und das ist das, worauf es hier ankommt – verändert sich durch die Unterstellung eines Eigenlebens die Beziehung des Menschen zur konkreten Technik, die dann leicht als echtes Gegenüber missverstanden werden kann, beispielsweise indem sie anthropomorphisiert wird (Kunkel 2023; van Oorschot 2023). Mit diesem Bild von Technik als echtem Gegenüber, als autonomen Selbst, kann sich der Mensch identifizieren, vergleichen und sich von daher auch verstehen. Und es liegt nahe sich vorzustellen, dass eine solche anthropomorph gestaltete Maschine auch als ein Sinnbild für den Wert menschlichen Lebens steht – auch wenn sie diesen nur simuliert (Coeckelbergh 2010, 237). Da es sich dabei jedoch um einen Trugschluss handelt, lautet die entscheidende anthropologische Frage welche Alternativen der Konstitution des Menschen in theologisch-ethischer Perspektive zur Verfügung stehen, wie der Mensch sich also besser verstanden weiß als durch die Reflexion, die Spiegelung mit Hilfe von Technik.

In der christlichen Tradition ist hierbei die Vorstellung zentral, dass der Mensch nach Gen 1,26 als Ebenbild Gottes, als sein צלם (*zelem*), erschaffen wurde. Die Interpretation dieses theologischen Terminus ist keineswegs einfach und eindeutig, sondern hochkomplex und vielschichtig (Neumann-Gorsolke

2017; Etzelmüller 2021, 278).[3] Die neuere alttestamentliche Forschung kommt jedoch überein, dass es sich bei dem Begriff am ehesten um eine Bildrepräsentation handelt, ähnlich wie in der alttestamentlichen Vorstellungswelt und seiner Umwelt der weltliche König Gott repräsentiert (Neumann-Gorsolke 2017; Etzelmüller 2021, 278f). »Der Mensch soll, wie der ideale König nach Psalm 72, für Recht und Gerechtigkeit sorgen – und so dazu beitragen, dass Gottes gute Intentionen für seine Schöpfung realisiert werden.« (Etzelmüller 2021, 279) Gemeint ist damit eine Beziehungsaussage: Es geht um das Verhältnis zwischen Mensch und Mit-Welt, um die sorgende Beziehung des Menschen zu seiner Umwelt, sei es in Form anderer Menschen, oder der Natur. Ähnlich wie es im alten Orient zu den Aufgaben des Königs zählte, für einen umfassenden Rechtszustand zu sorgen, so wird in der Rede vom Ebenbild die Aufgabe eines jeden Menschen ausgedrückt einen solchen Zustand handelnd anzustreben. Recht und Gerechtigkeit bilden das Zentrum menschlichen Handelns und Drücken das Ziel menschlicher Bestrebungen aus. Insofern es sich bei dieser Interpretation nun um ein Strebeziel in relationalen Verhältnissen der Menschen untereinander handelt – und weniger um eine ontologische Aussage – ist damit zugleich ein Sollen ausgedrückt: Es handelt sich also um eine Aufgabe, die dem Menschen gestellt ist. Und in dieser relational erwachsenen Aufgabe konstituiert sich der Mensch als Mensch in Beziehung zu anderen Menschen – und zwar aus der ursprünglichen Relation zu Gott heraus. Dadurch bleibt der Mensch gerade nicht in seiner Relation zu sich selbst, sondern überschreitet diese mit Blick auf Gott. Dies wiederum kann als Korrektiv auch für die Beziehung zur Maschine hilfreich sein, denn so weitet sich der Blick des Menschen auf die ganze Fülle anthropologischer Bezüge: Eigenschaften, die sich mit Recht und Gerechtigkeit verbinden, etwa. In anderen Worten: Eine Beziehung, in der ein Mensch sich vor den Rechenoperationen der Maschine schämt und infolgedessen bereit ist, die damit verbunden Handlungen nicht mehr selbst auszuführen, übersieht die Defizite der Maschine. Er gibt damit nicht nur die Rechenoperationen aus der Hand, er versteht sich selbst nicht

3 Das betrifft beispielsweise die zentrale Frage inwiefern die alttestamentlichen Aussagen der Gottesebenbildlichkeit aller Menschen in Spannung stehen zum neutestamentlichen Befund, dass allein Jesus Christus eben dieses Ebenbild darstellt (Peters 2010). Oder aber, ob es sich bei der Ebenbildlichkeit des Menschen um eine ontologische Qualität handelt, oder eine relationale Aussage, die das Welt- bzw. Gottesverhältnis des Menschen thematisiert (Neumann-Gorsolke 2017). Ich werde mich im Folgenden auf jene Themen und Fragestellungen konzentrieren, die für die Betrachtung meiner Fragestellung relevant sind.

mehr länger als jemanden, der in der Lage ist diese Operationen selbst auszuführen.

Entscheidend für meine Ausführungen hier ist aber, dass sich mit der Konstitution des Menschen von Gott her zugleich Perspektive und Ausrichtung verändern: Als Ebenbild Gottes ist der Mensch an Gott orientiert: er weiß sich geschaffen und findet Antwort auf die Frage des Woher. Er ist von vornherein als Beziehungswesen ausgewiesen. Und als solches an Gott orientiertes Beziehungswesen (Etzelmüller 2021, 283) kommen zugleich wesentliche Beziehungseigenschaften in den Blick wie Liebe, Mitleid oder Erbarmen. All dies sind Eigenschaften, die das Verhältnis zum anderen herausstellen und aufzeigen, dass erst in einer gelungenen Beziehung der Mensch zum Menschen werden kann (Etzelmüller 2021, 283). Richtet der Mensch nun aber seinen Blick auf die Technik, schaut er letztlich nicht auf diese Beziehungseigenschaften, sondern nur auf bestimmte kognitive Bereiche, Rechnen etwa, in denen die Technik dem Menschen überlegen ist.

Das darf nun aber nicht missverstanden werden. Mit der symbolischen Rede vom Ebenbild Gottes geht es mir nicht darum zu kritisieren, dass der Mensch sich selbst mit Gott verwechselt und sich darin ein Götzenbild schafft – so kann ja die Vorstellung der prometheischen Scham auch gedeutet werden: Indem Menschen sich an der Technik orientieren, schaffen sie sich selbst einen Gott, weil sie sich an technischen Vorstellungen ausrichten.[4] Ich verstehe aber Anders' Hinweis auf die prometheische Scham eher so, dass der Mensch sich in der Technik wiedererkennt und sie zu einem inneren Anderen wird, an dem er sich orientiert und ausrichtet, indem er sich vor ihm schämt. Diese Konstruktion des Anderen jedoch hebt nur bestimmte Eigenschaften des Menschen hervor und übersieht dabei zugleich blinde Flecken der Technik. Konkret heißt das, dass eine Technik, die Muster mit großer Treffsicherheit erkennt, etwa bestimmte Menschen oder Tiere, dadurch noch lange nicht weiß, was es heißt ein Mensch zu sein und wie ein Mensch zu fühlen. Also ist sie auch nicht in der Lage *Mit*-Gefühl oder *Mit*-Leid zu haben, das sich dann in erbarmungsvollem Handeln niederschlägt. Genau dies – so zumindest mein Vorschlag – versinnbildlicht das Symbol des *imago dei*. All diese Eigenschaften kann aber die Maschine qua Konstruktion nicht haben, denn es fehlt ihr nicht nur das dafür nötige Weltwissen, sondern auch das Selbst, das Bewusstsein

4 Dies etwa kritisiert Noreen Herzfeld in ihrem Artikel In Whose Image? Artifical Intelligence and the Imago Dei (Herzfeld 2012). Ähnlich verstanden werden kann auch das Buch Homo Deus von Yuval Noah Harari (Harari 2017).

(Smith 2019, 97–103). Die damit verbundene Gefahr jedoch, geht darüber hinaus, nämlich indem der Mensch sich vorwiegend oder nur noch vom Bild der Maschine her versteht – und damit letztlich Eigenschaften wir Mitleid und Erbarmen nicht mehr in sein Repertoire aufnimmt. Dass solche Gedanken nicht ganz von der Hand zu weisen sind zeigt sich da, wo Robotiker*innen wie Ronald C. Arkin dafür argumentieren, dass die Technik im Schlachtfeld letztlich menschlicher als der Mensch operieren kann, weil ihr eben jene menschlichen Eigenschaften, bzw. Unzulänglichkeiten fehlen. In diesem Bild, klingt deutlich die anfangs zitierte Aussage von Sarah Connors an: Mitmenschlichkeit können – und sollten – wir von der Maschine lernen, weil diese Normen der Mitmenschlichkeit weitaus besser umsetzen kann als der Mensch selbst. Dass sich diese Argumentation letztlich selbst widerspricht, zeigt sich schon allein daran, dass ein autoregulatives Waffensystem – oder im Falle Connors der Terminator – schlicht und einfach kein Mensch, sondern eine Maschine ist und dass allein deswegen die Vorstellung gerade von einer Maschine, die zum Töten programmiert wurde, Mitmenschlichkeit zu lernen absurd ist. In anderen Worten: Wer Mitmenschlichkeit von einem autoregulativen Waffensystem erwartet, hat einen Kategorienfehler begangen. Die soziale Imagination, die hinter diesen moralischen Vorstellungen steht, ist mit den technischen Voraussetzungen schlechterdings nicht vereinbar.

5. Zusammenfassung

Ich habe hier vorgeschlagen, die gegenläufigen Narrative des *Killerroboters*, bzw. des Waffensystems, das mitmenschlicher als der Mensch agiert, mit Hilfe von Charles Taylors Theorie der sozialen Imagination zu analysieren. Dabei zeigt sich, dass die jeweils erzeugten Bildwelten mit der Idee übermenschlicher Technik agieren, die dann also solche wiederum Rückwirkungen auf das Menschenbild hat, da der Mensch sich von dieser Technik her (miss-)versteht.

Ich schlage dagegen vor mit dem religiösen Symbol des *zelem adonai* die Konstitution des Menschen als mitmenschliches Beziehungswesen zu verankern. Denn für Gott – und entsprechend den Menschen – scheint mir weder die Fähigkeit zur Mustererkennung noch zum Rechnen entscheidend, sondern die zum Mitleiden und Erbarmen, zum Handeln und Fühlen. Und eben diese Fähigkeiten finden sich in der Maschine nicht. Das Fazit lautet dann, dass der Mensch sich über die Eigenschaften des Fühlens und Mitleidens, des Handeln und Erbarmens verstehen sollte, weil diese grundlegend für das Weltverhältnis

des Menschen sind. Der ethische Imperativ, der sich damit zugleich verbindet, ist, dass eben diese menschlichen Eigenschaften nicht der Maschine anheimgestellt werden sollten: Rechnen soll also die Maschine, mitfühlend handeln aber der Mensch. Das heißt aber auch, dass Prozesse, die mitfühlendes Handeln erfordern, der Maschine nicht überlassen werden dürfen.

Um schließlich Sarah Connor noch einmal deutlich zu widersprechen: Mitmenschlichkeit von einem Terminator lernen zu wollen ist ein Ding der Unmöglichkeit! *Mit*-menschlichkeit zu lernen ist nur möglich in Auseinandersetzung und im Angesicht von Menschen. Und um zu lernen, was der Mensch ist, so lautet zumindest mein Vorschlag, lohnt sich ein Blick in die christlich-religiöse Symbolwelt, die mit Bildern wie *Ebenbild Gottes* die besonderen Eigenschaften des Menschen als Beziehungswesen, als *Mit*-Mensch herausstellt.

Literatur

Anders, Günther. 1956. *Die Antiquiertheit des Menschen: Über die Seele im Zeitalter der zweiten industriellen Revolution*. München: C.H. Beck.

Arkin, Ronald C. 2009. *Governing lethal behavior in autonomous robots*. Boca Raton: CRC Press.

Campaign to stop killer robots. n.d. »Less autonomy: more humanity.« Accessed April 19, 2023. https://www.stopkillerrobots.org/.

Coeckelbergh, Mark. 2010. »Moral appearances: emotions, robots, and human morality« *Ethics Inf Technol* 12: 235–241.

Coeckelbergh, Mark. 2017. *Using words and things. Language and philosophy of technology*. New York/London: Routledge.

Etzelmüller, Gregor. 2021. *Gottes verkörpertes Ebenbild: eine theologische Anthropologie*. Tübingen: Mohr-Siebeck.

Fuchs, Thomas. 2020. »Menschliche und Künstliche Intelligenz: Eine Klarstellung.« In *Verteidigung des Menschen. Grundfragen einer verkörperten Anthropologie*, edited by Thomas Fuchs. Berlin: Springer, 21–70.

Geiß, Robin, Henning Lahmann. 2018. »Autonomous weapons systems: A paradigm shift for the law of armed conflict?« In *Research handbook on remote warfare*, edited by Jens David Ohlin. Cheltenham/Northampton, 371–404.

Harari, Yuval Noah. 2017. *Homo Deus: A brief history of tomorrow*. London: Vintage.

Herzfeld, Noreen. 2012. »In Whose Image? Artifical Intelligence and the Imago Dei.« In *The Blackwell Companion to Science and Christianity*, edited by J.B. Stump, Alan G. Padgett. Hoboken: Wiley, 500–509.

Höhne, Florian. 2022. »Bilder des Menschlichen: Theologisch-ethische Herausforderungen der Vorstellungswelten künstlicher Intelligenz.« In *Framing KI: Narrative, Metaphern und Frames in Debatten über Künstliche Intelligenz*, edited by Frederike van Oorschot, Selina Fucker. Heidelberg: heiBOOKS, 111–136.

Koch, Bernhard. 2016. »Maschinen, die uns von uns selbst entfremden. Philosophische und ethische Anmerkungen zur gegenwärtigen Debatte um autonome Waffensysteme.« *Militärseelsorge: Dokumentation* 54: 99–119.

Kunkel, Nicole. 2021. »Autoregulative Waffensysteme: Automatisierung als friedensethische Herausforderung – ein Werkstattbericht.« *Ethik und Gesellschaft* 2. DOI: https://dx.doi.org/10.18156/eug-2-2021-art-6.

Kunkel, Nicole. 2023. »Programmierte Autonomie? Autoregulative Waffensysteme als anthropologische Anfrage.« In *Digitale Transformation der Gesellschaft. Neubestimmung des Sozialen durch Technik*, edited by Sebastian Kistler, Anna Puzio, Anna-Maria Riedl, Werner Veith. Münster: Aschendorff.

Müller, Vincent C. 2016. »Autonomous killer robots are probably good news.« In *Drones and responsibility: Legal, philosophical and sociotechnical perspectives on the use of remotely controlled weapons*, edited by Ezio Di Nucci, Filippo Santoni de Sio. London: Ashgate, 67–81.

Neumann-Gorsolke, Ute. 2017. »Gottebenbildlichkeit (AT)« In *Das Wissenschaftliche Bibellexikon im Internet*. Accessed April 18, 2023. https://www.bibelwissenschaft.de/stichwort/19892/.

Peters, Albrecht. 2010. »Bild Gottes: IV Dogmatisch.« *TRE*, 6, 506–517.

Ronald C. Arkin. 2014. »Vollautonome letale Waffensysteme und Kollateralopfer.« *Ethik und Militär* 1: 3–12.

Scharre, Paul. 2016. *Autonomous weapons systems and operational risk: Ethical autonomy project 2016*. Accessed April 10, 2023. https://www.cnas.org/publications/reports/autonomous-weapons-and-operational-risk.

Smith, Brian Cantwell. 2019. *The Promise of Artificial Intelligence. Reckoning and Judgement*. Cambridge & London: MIT Press.

Taylor, Charles. 1992. *Sources of the Self: The Making of the Modern Identity*. Cambridge: Harvard University Press.

Taylor, Charles. 2004. *Modern Social Imaginaries*. Durham & London: Duke University Press.

Terminator 2. 1991. »Judgement Day«. Accessed April 18, 2023. https://scripts-onscreen.com/movie/terminator-2-judgement-day-script-links/.

The Future of Live institute. N.d. »Slaughterbots are already there.« Accessed April 10, 2023. https://futureoflife.org/project/lethal-autonomous-weapons-systems/.

Van Oorschot, Frederike. 2023. »Alles Technik oder was? Ethische Perspektiven auf das Verhältnis von Mensch und Maschine im Kontext einer imaginationssensiblen Technikethik.« In *Mensch und Maschine im Zeitalter »Künstlicher Intelligenz«: Theologische Herausforderungen*, edited by Hermann Diebel-Fischer, Nicole Kunkel, Julian Zeyher-Quattlender. Münster: LIT. [Im Erscheinen].

Democratic Autonomy vs. Algorithms? Limits and opportunities for public reasoning

Sophie Jörg

1. Autonomy and Technology: A glance at a long tradition

Thanks to the ever-growing availability of data and the increasing power of algorithms, political practices and processes are changing. But what are the implications of the increasing use of algorithms for people's political or democratic autonomy? Do algorithms have an *empowering* or *disempowering* effect on us citizens? Drawing on references from the philosophy of technology as well as the theory of democracy, this paper makes the case for a differentiated, functionality-based approach towards the impact of algorithms for public reasoning. By shedding light on the curation, moderation and verification of information by algorithms, it is argued that democratic autonomy – i.e., individuals' jointly reasoned self-rule or public reasoning – is empowered by algorithms if they improve the quality of discourse and utilize the process conditions of public reasoning as benchmarks for moderating content. Whereas algorithms disempower citizens if they are used as automated shortcuts to the presuppositional process of jointly reasoned self-rule of citizens.

The public debate about the social and political impact of aritifical intelligence (AI), machine learning- (ML), and deep learning- (DL) technologies in liberal democracies seems to be overtaking the earlier concern about the replacement of humans by machines. In this context, amongst others, it is discussed how digital infrastructure alters the process of democratic elections and public opinion-forming, the voter turnout as well as the voter behavior, or how DL can affect political communication in general. Against this backdrop, new theories of democracy – such as E-Democracy (e.g., Kneuer 2019),

Data Democracy (e.g., Batarseh/Yang 2020), or Liquid Democracy (e.g., Paulin 2020) – emerged.

Yet, all these issues – raised in political philosophy as well as political science – resonate with the fundamental question of both the humans' and the technology's *agency* and their respective *autonomy*: Do digital technologies in fact expand the scope of human action or do they restrict it? How does this affect democratic practices and their political legitimacy? By contrast, what kind of agency is ascribed to DL-technology, for example?

Although the debate about human autonomy has experienced a renaissance in the wake of technical automation, exploring the relation between technology and autonomy has a long tradition as it has always been a key focus of machine ethics and philosophy of technology (Heßler 2019). Autonomy, here, is essentially negotiated on both levels: The autonomy of humans on the one hand and the autonomy of technology on the other; resulting in a twofold meaning of the concept: human autonomy and technological autonomy or ›autonomous systems‹ (cf. Chiodo 2022).

However, with the recent advances in DL and ML algorithms and their ubiquitous use, the long-standing research on the correlation between technology, autonomy, and automation has gained new momentum. The improvement of complex algorithms inscribed – or rather coded – technology with a higher degree of agency resembling in a decisive caesura regarding human-computer-interaction (HCI). To date, the effects of this caesura on human autonomy have been studied primarily at the individual level, along the lines of the enlightenment concept of self-determination (individual or personal autonomy) (cf. Laitinen/Sahlgren 2021; Sankaran et al. 2020). Whereas the systemic dimension of autonomy in the political sense of self-legislation (political or public autonomy), as coined especially by liberal theorists of democracy, is comparatively omitted – despite the prominent references to the potential threats Tech might pose for democracy (cf. Sætra 2021).

The aim of the paper is to further differentiate the analysis of human autonomy in the context of algorithm-based technology with regard to the democratic theoretical implications of the algorithm-driven shift in HCI. In doing so, the reflexive consideration of individual's personal autonomy shall be extended to the notion of autonomy as self-rule on group level, as Max Weber might say (Weber 2006 [1948]). To this end, human autonomy will be considered not only as an (intrinsic) end in itself, but as a necessary and functional element in the fabric of democratic practice. Against this background, the paper discusses both the drawbacks and merits of algorithms and algorithmic

decision-making in the public sphere for citizens realizing their democratic autonomy.

2. Autonomy and Algorithms: Conceptual references for an analysis

Everyday life is permeated by digital, algorithm-based search engines (e.g., Google, Bing), translation programs (e.g., Google Translator, DeepL), or recommendation systems (e.g., AboutYou, Netflix, Spotify) that serve as decision filters for various purposes. Algorithms thus not only influence our taste in clothes, movies, or music (Goldschmitt and Seaver 2019), what information we receive in our social media news feed (Pentenrieder 2021, 53), they might also determine our credit score or political attitude (Cho et al. 2020). »Algorithms«, according to Verständig et al. (2022, 8), »are invisible and yet ubiquitous and tangible«[1]. So, in some respect they might actually determine our thoughts and actions. But does this ›algorithmization‹ or ›platformization‹ of the public sphere and communication, as recently postulated by Habermas (2022) or Eisenegger (2021), also amount to a dilution of the legitimacy of political decision-making in liberal democracies as it infringes people's democratic autonomy? To further address this question, we will first review the conceptual roots of the two main concepts – Autonomy (2.1) and Algorithms (2.2) – in light of the aforementioned triad of automation, autonomy and democracy.

2.1 Democratic Autonomy as *jointly reasoned self-rule*

During the European Enlightment the philosophical concept of autonomy found its way into Western societies and has been regarded as a core concept of modernity ever since (Thimm/Bächle 2019, 75). As an antonym to external determination, it originates in Greek (›*auto*‹ = *self*; ›*nomos*‹ = *law*) and refers to an individual's self-determination or self-rule through which people can participate as subjects on their own authority (Quante 2013, 47). With Kant, the term was linked in the 18th century to the use of human reason, which affected the discourse about supposedly ›autonomous technology‹ to this day (Heßler 2019, 249). With the prominent saying ›*sapere aude*!‹ (*engl*. ›Have courage to seize your own reason!‹), Kant's philosophy paved the way out of mankind

1 Translated by the author.

from ›self-inflicted immaturity‹. By means of the reason bestowed upon them, humanity could free itself from paternalism and foreign domination. Personal (or individual) autonomy in this context means the self-determination of the human subject through reason. Consequently, autonomous subjects are those who have their own freedom of will and decision by utilizing their ability to think critically and revise their stances respectively (Pohlmann 2010).

In the 20th century, theorists, such as Habermas (1992) and Rawls (1996), further differentiated the notion of autonomy in terms of its political dimension (cf. Thimm/Bächle 2019, 76). Rawls, for example, conceptualized autonomy – similar to ancient Greece – as a political category for the preservation of citizens' freedom. Unlike personal autonomy, this type of autonomy first appears in the political sphere or the political activity of citizens (Weithmann 2011, 327). According to Rawls, it would be realized by citizens »[…] participating in society's public affairs and sharing in its collective self-determination over time« (Ibid.). Habermas, however, argues that this shade of civic or public autonomy should not be considered separately from the private or personal autonomy described at the outset. On the contrary, both concepts are co-original (*gleichursprünglich*), mutually dependent and equally fundamental (Habermas, 2019, pp. 134, 163; Jörg 2023, 4): Public autonomy, in a nutshell, presupposes private autonomy, because it requires a legal system that is legitimate only if it guarantees equal freedoms to its subjects. Private autonomy, on the other hand, requires public autonomy, as the legal regulation of private autonomy is legitimate only if it emerges from a discursive process that guarantees political rights.

In regard to the emphasis on civic self-legislation, Habermas's understanding of autonomy is described in the literature as an »attempt to rethink the Kantian (and Rousseauian) idea of individual freedom through self-governance« (Anderson 2019, 22). The concrete exercise of self-governance, then, takes place within the framework of the public, discursive formation of opinion and will (Habermas 2019, 161). Accordingly, political autonomy can only be achieved by »equal opportunities to participate in processes of opinion- and will-forming«[2] (Habermas 2019, 156). In this respect, the entangled concepts of public and private autonomy are essential for a functioning democracy, as they pinpoint the necessary capacity of humans to (self-)reflect, critically assess, and soundly evaluate solutions regarding complex political, social, or individual conflicts (cf. Richardson 2002).

2 Translated by the author.

These philosophical references, however, inform the conception of the so-called democratic autonomy, which in essence describes the principle of people's jointly reasoned self-rule. In this notion of human autonomy one of the most central – if not the most central – democratic maxim culminates: the primal demand that »we [the citizens] must reason together in order to rule ourselves« (Richardson 2002, 18). Realizing citizens' democratic autonomy, hence, is constitutive for the democratic constitution of political order and its legitimacy. Drawing on the co-originality of public and private autonomy, then, the idea of being »the author of my private life« (*personal autonomy*) is closely tied to the idea of »joint authorship of our common political life« (*public autonomy*) (Lovett/Zuehl 2022, 469). With this merger of individual self-determination and systematic self-rule or self-governance both influential manifestations of autonomy are conceived together and united in the term's core as ›public reasoning about the ends of policy‹ (Richardson 2002). Drawing on Habermas' influential theory of democracy, this public reasoning, in turn, is closely linked to four crucial – albeit demanding – preconditions: The (i) *inclusivity* of all possible topics and participants, (ii) *equal* distribution of *participation* opportunities, the (iii) *sincerity* of all participants, and finally the (iv) *absence* of a *coercive communication structure* (Habermas 2009, 89). Accordingly, democratic autonomy or jointly reasoned self-rule manifests itself in the process of public reasoning; meaning the democratic process of public opinion-forming and the public decision-making based on it.

2.2 Algorithms as mechanism of technical *automation*

Algorithms are often recognized as the technical underpinning for technology's alleged ›agency‹ and ›autonomy‹. In the following, however, it is argued that they equal mechanisms of pure technical automation, which are often misinterpreted as autonomous, i.e., self-determined, action of technology. What may initially appear to be autonomous action in the course of technical processes actually turns out to be nothing more than a complex automation of various computing steps programmed by algorithms in order to perform specifically defined tasks. While exploring the automation of technology by algorithms, possible limits of a ›technological autonomy‹ are briefly pointed out.

However, to understand why exploring algorithms is crucial to be aware of the implications of digital technologies for democratic legitimacy of political decision-making, it is worth taking a glance at the technical aspects of

algorithms: In computer science an algorithm is acknowledged to be a »list set of instructions, used to solve problems or perform tasks, based on the understanding of available alternatives« (IIG 2023). Algorithms are systematic and mechanical methods for solving a well-defined problem, such as efficiently searching or sorting vast amounts of data (Belford/Tucker 2023). Due to detailed »specifications for performing calculations, data processing, automated reasoning or decision making« algorithms in fact automate technology's computing (IIG 2023). To summarize, computerized algorithms are »structured decision-making processes that *automate* computational procedures to generate decisional outcomes on the basis of data inputs« (Gal 2018, 64; emphasis by author). They are the technical foundation for many modern technologies that we use every day. Digital platforms, such as Google, Netflix or Instagram, are just a few examples of this.

In this sense, digital, algorithm-based technologies, such as AI, could initially be ascribed a certain degree of agency as these systems can partly perform the task assigned to them independently or without further human input. However, a glance at the interrelation of agency and autonomy reveals that those processes which *prima facie* may appear to be a kind of autonomous action of the technology turn out to be merely a form of precise – albeit by now highly complex – automation. Within the philosophical discourse it is widely acknowledged that an entity's ›agency‹ manifests its capacity to act or exert power on its own (Schlosser 2019). Akin to the concept of ›autonomy‹, it denotes a comparative state of mind – i.e., a state that can be gradually increased or decreased. Depending on a person's or system's agency – meaning their ability »to act on the beliefs and values they hold« (Prunkl 2022, 3) – these agents can be considered more or less autonomous or self-determined. Hence, both philosophical concepts agency as well as autonomy are intertwined. Yet, their relation indicates that a sufficient degree of agency enables autonomy in the first place. Against this background, it can therefore be presumed that algorithm-based technologies – to a certain extent – operate as actors. Yet, this does not inevitably imply that they are autonomous, since the computational processes or actions of AI, for example, are based on previously defined decision paths and not on the rationality or free will of the technology in question. In this sense, the technology lacks its very own impetus and reasoned deliberation to perform an action in question. Hence, algorithm-based systems do not operate out of intrinsic or extrinsic motivations, but because they have been created by humans to process these tasks by means of specialized algorithms. Although algorithms can ultimately provide certain recommendations

– e.g., which movie might be of interest to us or whether a melanoma is good or malignant – such ›decisions‹ or ›actions‹ are, hence, not based on the free will or the rationally guided deliberation of the processing technical system. Rather, the machine merely executes a sequence of programmed, previously well-defined computational steps that have been inscribed into the technology by computer scientists. The »decisional parameters and rules for weighing them«, however, are originally »set by the algorithm's designer« (Gal 2018, 64f.). So, algorithm-based technology can never be fully autonomous. Or as Russell, a leading AI researcher and professor of computer science at the University of California (Berkeley), stated: »[…] machines, unlike humans, have no objectives of their own, we give them objectives to achieve. In other words, we build machines, feed objectives into them, and off they go. The more intelligent the machine, the more likely it is to complete that objective« (Russell 2019).

So concludingly, computer scientists create the functional spectrum of mechanical autonomization by programming tailored algorithms, which in turn have enormous effects or backlashes on us. Regarding the rapid developments in the field of algorithm-based technology, it is therefore imperative to further question how algorithm-driven automation affects us citizens in our democratic autonomy.

3. ›Algorithmization‹ of Political Communication: Implications for public opinion-forming and decision-making

In the following, these backlashes will be systematically assessed regarding the above-mentioned conditions of joint public reasoning according to Habermas. Central to this, however, are less the informatics underlying algorithms than their political effects and objectives: What functions, for example, do algorithms fulfill in the context of political communication? Do they serve as a decision guidance for human action, as they are often portrayed (e.g., Liang et al. 2022), or do they gradually disenfranchise humans in the context of their political decision-making?

The subtle ›algorithmization‹ of political communication can be witnessed in various settings: Political actors – such as elected officials or representatives of interest groups and NGOs – for example, use technological proxies in the form algorithmically (semi-)automated chat bots to disseminate their messages in the most cost- and time-efficient way (Woolley/Howard 2016). Moreover, algorithms are utilized to provide people with different messages.

By means of so-called micro-targeting strategies, people are given different information depending on their algorithmically calculated political interests, demographic and social profiles (Zarouali et al. 2022). Further, DL-algorithms are used to create or replace the image of one person in videos and other digital media, for example a high-ranking member of parliament or a military. So-called deep fakes are generated thanks to advanced algorithms. So, algorithms can be utilized for strategically controlling political communication and thereby political opinion making by the distortion of the supposed public opinion in various ways. Acting as a digital, stochastic calculating *spin doctor*, they deeply interfere in the process of shaping political opinion. Capable of shattering trust in politics and the overall facticity of information, they ought not to be underestimated (Whyte 2020). Algorithms do not only (pre-)select »what information is considered most relevant to us« (Gillespie 2014, 167) and then decide with whom which information shall be shared, they also generate and curate information by themselves. Mostly neither with our knowledge nor consent, algorithms manage our interactions on social networking platforms, our personal preferences, and whether we participate in a political discourse or not. In other words, algorithms do not just alter political communication but create the digital public sphere around us, in which we can exchange ideas with others. The amphitheater-like structure of the public sphere, in which politically relevant issues are debated and solutions sought together, however, is giving way to algorithmically computed filter bubbles and echo chambers – i.e., parallel ›public spheres‹ in which people no longer exchange views with dissidents but with like-minded people. This algorithm-driven fragmentation of the public sphere and the resulting impediment to public deliberation are only some – yet the most prominent – destructive consequences Habermas analyzes in his latest work *A New Structural Transformation of the Public Sphere and Deliberative Politics* (2022).

With this power of governing the flows of information, algorithms change political communication drastically. By shaping communication channels and their respective content, algorithms deeply intervene in the process of public opinion- and decision-making. In this sense, the algorithmization of political communication equates to a radical intrusion into democratic processes in general and citizens' ability to truly exercise their democratic autonomy in particular; as the state of information and the algorithmically induced fragmentation of the public sphere have a very significant influence on how and whether people can and want to engage in discourse: algorithms inevitably impact the *accessibility* and *inclusivity* of discourse, the *equality* of contributions and oppor-

tunities for participation, the *sincerity* of all participants, and the *informality* of exchange. Due to algorithms, the supposedly invisible becomes visible and the supposedly visible becomes invisible. What may feed new, rather minority voices and ideas into the discourse, may also cancel out opinions or deliberately bring them into focus.

To conduct a more nuanced analysis of the algorithm-driven expansion or restriction of democratic autonomy and action, it seems useful to first take a more nuanced look at algorithms' various functions in dealing with information. Drawing on Sauerwein et al. (2022) research, three key functions can be distinguished: the (I) curation and filtering of existing content, the (II) moderation of specific content, and finally the (III) verification of this content. In the following, the effects of all these categories on people's democratic autonomy or their jointly reasoned self-rule will be considered.

3.1 Curation and filtering of content

The curation and filtering of information – i.e., finding, compiling and organizing relevant content – is one of the most discussed functions of algorithms in the context of political communication (cf. Berman/Katona 2020). According to Berman and Katona (2020, 298f.), the literature on curation algorithms initially focused on the economic design and the impact of algorithmic curation and filtering on consumers. Other researchers examined the impact on user consumption behavior and any change in terms of the streamed information's quality (Athey et al. 2017). The citizen perspective, on the other hand, lately received attention with studies on trends of polarization of discourse triggered by curating algorithms, increasing authorship by users or more active participation in online discussions, and the spread of political ideologies through pre-filtering algorithms (Cho et al. 2020).

However, the observed and expected effects of this seem to be ambivalent. At first, the potential of algorithmic curation gives rise to hopes of a substantial expansion of democratic autonomy: By reliably processing the large amount of information and data available today according to appropriate criteria, information that is relevant for discourse or common reasoning can be identified, filtered, and prepared for further reflection. By means of their advanced and robust pattern recognition, algorithms could not only distinguish valid contributions from illegitimate ones – for example, hate speech and other defamation and thus, according to Habermas, illegitimate content in political discourse – but contributions could also be pre-structured

into specific argumentation patterns or core statements. In this way, factual arguments could be filtered out of vast piles of data and fed into the discourse in a way that is prepared for further discussion. In addition to possible strands of argumentation, it would also be possible to cluster interest groups and the premises or demands underlying the contributions, which could be hierarchized in a further step by means of algorithms according to their logic and stochastically mapped according to their agreement in the discourse. Such possibilities would not only make it easier for citizens to obtain relevant information more quickly, but also provide them with an already sensibly prefabricated basis for the further reflection process. Curating and filtering information through algorithms can not only simplify and optimize access to content relevant for legitimate, public opinion and decision-making, it also facilitates linking and confronting the participants and their ideas within discourse – regardless of their intellectual capabilities, individual language skills or eloquence etc. Advanced algorithms in natural language models, for example, are already capable of reading out written text, transforming spoken words into text and/or translating it into a desired language. Such accessibility, in turn, would allow more people – especially minorities, such as blind or illiterate people – to participate equally in political discourse and make their voices heard. Citizen's jointly reasoned self-rule and the public reasoning as such, it could be argued, would be simplified by the algorithm-based automation of these curations; the exercise of the democratic autonomy of citizens would thus be facilitated.

And yet, automating the curation of information by delegating it to algorithms does not necessarily lead to greater inclusion, participation, and a higher quality of political discourse. After all, the curation of information is always necessarily connected with the evaluation, selection or filtering of information, from which some pitfalls for democratic autonomy can be derived: Saurwein et al. (2022), for example, identify several risks associated with technically automated filtering and curation. For example, the emergence of errors and unwanted selections by algorithms. If, for example, admissible contributions were filtered in advance and excluded from the discourse based on incorrectly placed markers or keywords, this would not only violate the process condition of inclusion and accessibility, but also the participation rights of citizens. In this respect, their autonomy would be unjustifiably restricted by automation via algorithms. The targeted manipulation of citizens during the public opinion and decision-making process can also be seen as a significant restriction of democratic autonomy. This is because the parameters according

to which the responsible algorithms classify content as relevant or irrelevant are usually unknown to us. This intransparency or opacity of algorithm-based filtering and preselection of information ultimately leads to a downright distortion of public opinion formation – resulting in cutting citizens' self-determination. Citizens could no longer reason together, on an equal footing, when they have such a variety of information at their disposal. According to the current logic of algorithms, information flows are strongly tailored to individual needs and preferences. This logic is inevitably followed by information asymmetry, which makes reasoning on an equal footing, without constraints almost impossible. The emergence of such information and power asymmetries, on the other hand, can be additionally reinforced by strategies such as micro-targeting already mentioned.

By consciously and specifically controlling the flow of information through micro-targeting – i.e., addressing tailor-made messages, which is geared toward addressing pursuable or mobilizable citizens (Kruschinski/Haller 2017, 3) – information can be disseminated without passing through the public space and made accessible to all. This unequal or imbalanced exchange of information, in turn, can lead to a distortion of reality and thus of perceived public opinion. So political microtargeting by means of algorithms, however, does not necessarily lead exclusively to a reduction of democratic autonomy. Citizens can also be encouraged to participate in discourse when information is curated and filtered accurately and correctly. Tailored messages might be more relevant for users, helping them to better keep up with central political issues and arguments, which might amplify the effects of better voter turnouts and a more active citizenry exercising their democratic autonomy. In addition to this potential for mobilizing citizens, their political knowledge and general informedness could also be improved through algorithm-driven subject-oriented curation, which in turn could improve the fundamental quality of shared reasoning.

3.2 Moderation of specific content

As Gorwa et al. (2020, 1) state »[a]utomated hash-matching and predictive machine learning tools – what we define […] as algorithmic moderation systems – are increasingly being deployed to conduct content moderation at scale […] for user-generated content«. By means of these algorithmic moderation tools, problematic, toxic, or hostile statements such as hate speech can effectively be identified, deleted and prevented. With the increased use of algorithms for

content moderation, a growing body of research examines the political, social, and economic impact of this function of algorithms (cf. Kaye 2019).

Drawing on Gorwa et al. (2020, 3), we can distinguish between two moderating systems: Those, who decide about the visibility of accounts and their content (hard moderation) and systems utilizing rather soft moderation-techniques, such as recommendation systems or nudging by design. The political power of algorithms, however, becomes particularly evident in its function of hard content moderation: through the moderation of content – i.e., the technically automated decision-making as to which information is permitted in the discourse and which statements must be deleted or censored – a wide-reaching power but also responsibility is transferred from the people to the technology they design. Similar to curating, this second function of algorithms decides how and what information we perceive and what will be included into the political discourse. Against this background, algorithms also function as a new mechanism of gatekeeping. What traditionally has been performed by humans now is taken over by algorithms. Accordingly, algorithms are in charge of the content that is »visible and therefore noticeable to users, and influence the diversity of content that is consumed« (Stark et al. 2020, 10).

The use of these algorithm-based, automated techniques for hard and soft content moderation, on one hand, can potentially improve the level of discourse and in this way to support the legitimacy of political opinion and decision-making. Because algorithms can effectively contribute to supporting the processes of the jointly reasoned self-rule by allowing or excluding content according to the discourse-ethical conditions of these processes. Thereby, they create a democratic or ›safe space‹ in which everyone can freely express their respectful and legitimate opinion. In this way, it might be possible to mobilize citizens to join the discourse or public reasoning which feared the hostilities and hatred within unmoderated communication. In principle, they undermine the authorship of the people who have written hate comments and other illegitimate contributions, but in terms of democratic theory, those statements were not permitted in political discourse or were part of democratic autonomy anyways. On the other hand, gatekeeping by algorithms also restricts us in our autonomy, ability to act and freedom of decision, as they restrict the information and options through their moderation and gatekeeping and thus define the framework for our actions.

However, if the criteria according to which algorithms moderate content correspond to the commercial logic of the platform economy (*clickbaiting*) – instead of the above-mentioned discourse ethical premises – algorithms will not

remove fake news and hate speech etc., but rather spread it throughout society and political dispute. Thereby, incorrectly coded moderation algorithms might undermine the »epistemic potential« of political discourse.

3.3 Verification of content

In the context of political communication, algorithms are also used to verify content. For this purpose, algorithms are developed for reviewing claims »with reference information in the form of facts in a knowledge base« (Huynh/Papotti 2019, 689). Hence, the ability, applicability, and performance of fact-checking algorithms of course is limited to the set of knowledge with which humans trained them and the pool of information which is available to them – without any human supervision or reinforcement.

In their function of verifying the objective accuracy, truth or factuality of information, algorithms again fulfill the role of a modern, digital gatekeeper that seems to be successively replacing the former control mechanisms, such as journalists and mass media. While algorithm-based platforms and search engines, such as Google or Bing, facilitate access to content and therefore expand human agency, algorithms specially trained to identify fake news, deep fakes and other false statements additionally ensure human access to verified information and knowledge. This, in turn, not only enables free accessibility to objective facts or information, but it also increases people's informed self-determination.

How dangerous disinformation and the targeted dissemination of false statements can be for the integrity and legitimacy of political processes has become something of a commonplace in discourse (Morgan 2018). Yet, disinformation campaigns not only polarize and radicalize political discourse, they affect personal autonomy *per se*. As Witzleb and Paterson (2020, 227) note, mis- as well as disinformation in its various forms is »harmful as it has the potential to disrupt our individual capacity for self-authorship and, as a consequence, our communal capacity for self-government«. False information can interfere with democratic autonomy, as both the sincerity and the honest solution orientation of the agents involved in political debates can no longer be ensured.

The implementation of algorithms to contain the spread of misinformation and disinformation and its threat for democratic society and citizens' democratic self-determination can therefore rather be seen as an autonomy supporting than a diminishing function.

4. Democratic Autonomy vs. Algorithms: Modelling the limitation and expansion of democratic autonomy

Based on the previous discussion of the algorithmization of political communication, finally, an attempt will be made to model the potential limitation and expansion of democratic autonomy by algorithms. To this end, the developed theses on the political implications of the (I) curation and filtering, the (II) moderation and (III) verification of content by algorithms will be systematized and summarized.

Considering the process conditions of joint deliberation and democratic autonomy, the elaborated merits and drawbacks can be presented as follows. The table is divided into the preconditions of democratic autonomy and the discourse conditions of joint reasoning, on one hand, as well as the functions of algorithms in the context of political communication, on the other. The markings indicate which conditions are endangered by which function:

Table 1: Synthesis of limits and opportunities for democratic autonomy due to algorithms

Condition **Function**	Accessibility & Inclusivity	Equality of Contributions & Participation	Sincerity of all Participants	Informality of Exchange
Curation & Filtering				X
Verification				
Moderation		X		X

In conclusion, it became evident that the individual functions of algorithms point to their power as new control mechanisms of the information flow and gatekeepers of discourse. The algorithmically controlled automation of information and communication, it was shown, creates the danger of »*shortcuts*«. The arduous process of jointly public reasoning – i.e., the active exercise of democratic autonomy, which is tied to rather demanding conditions – could possibly be circumvented or, in some aspects, even replaced by algorithms. Like many democratic theorists, Lafont – a former student of Habermas – warns of such cutbacks, as they have occurred throughout history (Lafont 2020). The participation of citizens and the free exercise of their self-determination should under no circumstances be undermined by supposedly efficient shortcuts. The »ideal of self-government« as vital, »participatory aspect of democracy« (Lafont 2020, 7), threatens to be undermined by taking supposedly easier or more efficient paths.

By contrast, the technical autonomization of fundamental, democratic processes, such as one's own walk to the ballot box or the struggle to find solutions to political issues, can be seen as such a shortcut. The algorithm-based processing of information necessary for reflective participation in public decision-making and opinion-forming could already be interpreted as the preliminary stage of such a development. Political alienation and the impediment of true democratic autonomy are the consequences of such shifts. Or as Lafont claims: »Democratic *participation* in decision-making is essential to prevent an *alienating disconnect* between the political decision to which citizens are subject and their political opinions and will« (Lafont 2020, 22f.). Algorithms, it had been shown, appear in both forms: On the one hand, they serve us as decision-making aids by facilitating access to objective facts and consensus-oriented input. They promote the inclusion of the other and in particular of those who think differently. Whereas, on the other hand, algorithms also force the exclusion of the other and of those who think differently by creating deep fakes or by moderating content in discourse in a click-oriented rather than a consensus-oriented manner. In this way, they no longer function as a decision-making aid in political interaction, but as a distorting mirror of public opinion and perception, which makes discussion at eye level and the free exercise of democratic self-determination by all almost impossible. The danger of manipulation and thus, to a certain extent, the ›disenfranchisement‹ of citizens grows with every use of algorithms and bots to deliberately change the climate of opinion. Citizens, with their individual opinions, rights and

political views, are increasingly relegated to the background. They are reduced to a number to be recorded statistically.

Algorithms, on the other hand, are to be commended if they improve the quality of public reasoning. For example, through the targeted verification of content, the identification and subsequent filtering of toxic statements and the clustering of certain strands of argument. Algorithms should therefore be used to improve the level of political debate and thus also protect the ideal of democratic self-determination. On the other hand, they should not be used to subtly and implicitly as well as explicitly force citizens to forego their right of political self-determination or even to no longer be able to exercise it. Rather, they should be used »*to force the political system to take ›the long road‹ of properly involving citizenry*« (Lafont 2020, 159).

5. Concluding Remarks

Thanks to the ever-growing availability of data and the increasing power of algorithms, political practices and processes are changing. But what are the implications of the increasing use of algorithms for people's political or democratic autonomy? Do algorithms have an empowering or disempowering effect on us citizens?

The paper aimed to address this question by drawing on references from the philosophy of technology as well as from the theory of democracy. In course of this, the philosophical concept of autonomy was extended to the political idea of so-called democratic autonomy, as elaborated by Habermas. The paper, further, offered a short note on the importance of algorithms as key mechanisms for technical autonomation. It became clear that algorithms have a profound impact on the way people discuss current affairs and engage with each other as well as politics.

In conclusion, the paper argues that algorithms can be a threat to citizens' autonomy and their right to informational self-determination at the individual level, but they can also be challenging to the discursive process of opinion- and will-forming at the structural level. In both cases, however, algorithms decisively determine the basis on which humans' political decisions are made or an action is taken. Regarding the systematization of algorithms' functions and their impact on the preconditions of jointly public reasoning, the paper found that the (3.1) curation and filtering by algorithms infringes the necessary informality of exchange. Whereas algorithms (3.2) verifying content are empower-

ing humans in their ability to participate in discourse or the joint public reasoning, as it contributes to the improvement of the discourse's quality and thus counteracts the chilling effects and the radicalization and polarization of the discourse. The (3.2) content moderation by algorithms, in contrast, might diminish both the informality of exchange as well as the equality of contributions and participation in discourse.

As the paper noted, human autonomy has become a theme across various guidelines and principles on the responsible use and design of algorithms. However, according to Sætra, Borgebund and Coeckelbergh (2022, 804), there is a tendency within AI research to adopt the concept of democracy and its associated norms merely as buzzwords – without referring to their further political and historical significance. The paper's objective has been to shed light on the political dimension of the various effects the growing power of algorithms hold for humans' political or democratic autonomy. Against this background, fundamental questions remain untouched, such as whether political discourse will be obsolete one day because the better argument or the better solution for political conflicts could be found algorithmically. Are algorithm-based automations, such as finding political solutions through mathematical calculations instead of the deliberative ideal of communicative negotiation, be permissible in terms of democratic theory or should they be rejected as destructive shortcuts? These and many other questions remain to be researched in the future.

References

Anderson, Joel. 2011. »Autonomy, Agency, and the Self.« In *Jürgen Habermas. Key Concepts*, edited by Barbara Fultner, 91–114. Durham: Acumen.

Athey, Susan, Markus M. Mobius, Jeno Pal. 2017. »The impact of aggregators on internet news consumption.« Working paper, Stanford University, Stanford, CA.

Batarseh, Feras A., Ruixin Yang. 2020. *Data Democracy: At the Nexus of Artificial Intelligence, Software Development, and Knowledge Engineering*. Academic Press.

Belford, Geneva G., Allen Tucker. 2023. »Computer Science.« *Encyclopedia Britannica*. Accessed April 23, 2023. https://www.britannica.com/science/computer-science/Algorithms-and-complexity.

Berman, Ron, Zsolt Katona. 2020. »Curation algorithms and filter bubbles in social networks.« *Marketing Science*, 39(2): 296–316.

Chiodo, Simona. 2022. »Human autonomy, technological automation (and reverse).« *AI & society* 37: 39–48.

Cho, Jaeho, Saifuddin Ahmed, Martin Hilbert, Billy Liu, Jonathan Luu. 2020. »Do search algorithms endanger democracy? An experimental investigation of algorithm effects on political polarization.« *Journal of Broadcasting & Electronic Media* 64 (2): 150–172.

Eisenegger, Mark. 2021. »Dritter, digitaler Strukturwandel der Öffentlichkeit als Folge der Plattformisierung.« In *Digitaler Strukturwandel der Öffentlichkeit: Historische Verortung, Modelle und Konsequenzen*, edited by Mark Eisenegger, Marlis Prinzing, Patrik Ettinger, Roger Blum. Springer VS. 17–39.

Gal, Michael S. 2018. »Algorithmic challenges to autonomous choice.« *Michigan Technology Law Review* 25: 59–104.

Gillespie, Tarleton. 2014. »The relevance of algorithms.« *Media technologies: Essays on communication, materiality, and society* 167 (2014): 167–193.

Goldschmitt, K. E., Nick Seaver. 2019. »Shaping the Stream: Techniques and Troubles of Algorithmic Recommendation.« In *The Cambridge Companion to Music in Digital Culture*, edited by Nicholas Cook, Monique M. Ingalls, David Trippett. Cambridge University Press, 63–81. DOI: https://doi.org/10.1017/9781316676639.006.

Gorwa, Robert, Reuben Binns, Christian Katzenbach. 2020. »Algorithmic content moderation: Technical and political challenges in the automation of platform governance.« *Big Data & Society* 7 (1). DOI: https://doi.org/10.1177/2053951719897945.

Habermas, Jürgen. 2009. *Zwischen Naturalismus und Religion. Philosophische Aufsätze*. Frankfurt a.M.: Suhrkamp.

Habermas, Jürgen. 2019 [1992]. *Faktizität und Geltung*. Frankfurt a.M.: Suhrkamp.

Habermas, Jürgen. 2022. *Ein neuer Strukturwandel der Öffentlichkeit und die deliberative Politik*. Berlin: Suhrkamp.

Heßler, Martina. 2019. »Technik und Autonomie.« In *Autonome Systeme und Arbeit. Perspektiven, Herausforderungen und Grenzen der Künstlichen Intelligenz in der Arbeitswelt*, edited by Hartmut Hirch-Kreinsen, Anemari Karačic. Bielefeld: Transcript, 247–274.

Huynh, Viet-Phi, Paolo Papotti. 2019. »A benchmark for fact checking algorithms built on knowledge bases.« *Proceedings of the 28th ACM Interna-*

tional Conference on Information and Knowledge Management, November 2019, 689–698.

International Institute in Geneva. 2023. *Understanding Algorithms in Computer Science*. Accessed April 23, 2023. https://www.iig.ch/en-en/blog/computer-science/algorithm-computer-science-definition-and-understanding.

Kaye, David. 2019. *Speech Police: The Global Struggle to Govern the Internet*. New York, NY: Columbia Global Reports.

Kruschinski, Simon, André Haller. 2017. »Restrictions on data-driven political micro-targeting in Germany.« *Internet Policy Review* 6 (4): 1–23. DOI: https://doi.org/10.14763/2017.4.780.

Kneuer, Marianne. 2019. »E-Democracy.« In *Handbuch Digitalisierung in Staat und Verwaltung*, edited by Tanja Klenk, Frank Nullmeier, Göttrik Wewer. Springer Fachmedien: Wiesbaden, 267–277.

Lafont, Cristina. 2020. *Democracy without shortcuts: A participatory conception of deliberative democracy*. Oxford University Press.

Laitinen, Arto, Otto Sahlgren. 2021. »AI systems and respect for human autonomy.« *Frontiers in artificial intelligence* 4: 1–14.

Liang, Garston, Jennifer F. Sloane, Christopher Donkin, Ben R. Newell. 2022. »Adapting to the algorithm: how accuracy comparisons promote the use of a decision aid.« *Cognitive research: principles and implications* 7 (1):14. DOI: https://doi.org/10.1186/s41235-022-00364-y.

Lovett, Adam, Jake Zuehl. 2022. »The Possibility of Democratic Autonomy.« *Philosophy & Public Affairs* 50 (4): 467–498.

Morgan, Susan. 2018. »Fake news, disinformation, manipulation and online tactics to undermine democracy.« *Journal of Cyber Policy* 3 (1): 39–43.

Paulin, Alois. 2020. »An overview of ten years of liquid democracy research.« *The 21st Annual International Conference on Digital Government Research* June 2020: 116–121.

Pentenrieder, Annelie. 2021. »Algorithmen erklärt euch.« In *In digitaler Gesellschaft*, edited by Kathrin Braun, Cordula Kropp. transcript Verlag, 53–69.

Pohlmann, Rosemarie. 2010: »Autonomie«. In *Historisches Wörterbuch der Philosophie*, edited by Joachim Ritter, Karlfried Gründer, Gottfried Gabriel. Berlin: Schwab.

Prunkl, Carina. 2022. »Human autonomy in the age of artificial intelligence.« *Nature Machine Intelligence* 4 (2): 99–101.

Quante, Michael. 2013. »Autonomie«. In *Lexikon Philosophie. Hundert Grundbegriffe*, edited by Stefan Jordan, Christina Nimtz. Stuttgart: Reclam, 47–48.

Rawls, John. 1996. *Political Liberalism*. New York: Columbia University Press.

Richardson, Henry S. 2003. *Democratic Autonomy: Public Reasoning about the Ends of Policy*. Oxford: University Press.

Russell, Stuart. 2019. »How to stop Superhuman A.I. before it stops us.« *New York Times*. Accessed April 24, 2023. https://www.nytimes.com/2019/10/08/opinion/artificial-intelligence.html.

Sætra, Henrik Skaug. 2021. »A typology of AI applications in politics.« In *Artificial Intelligence and Its Contexts: Security, Business and Governance*, edited by Anna Visvizi, Marek Bodziany. Springer VS, 27–43.

Sætra, Henrik Skaug, Harald Borgebund, Mark Coeckelbergh. 2022. »Avoid diluting democracy by algorithms.« *Nature Machine Intelligence* 4 (10): 804–806.

Saurwein, Florian, Charlotte Spencer-Smith, Jaro Krieger-Lamina, et al. 2022. »Social-Media-Algorithmen als Gefahr für Öffentlichkeit und Demokratie: Anwendungen, Risikoassemblagen und Verantwortungszuschreibungen.« In *Digitalisierung und die Zukunft der Demokratie. Beiträge aus der Technikfolgenabschätzung*, edited by Alexander Bogner, Michael Decker, Michael Nentwich, Constanze Scherz. Baden-Baden: Nomos, 243–256.

Sankaran, Supraja, Chao Zhang, Mathias Funk, et al. 2020. »Do I have a say? Using conversational agents to re-imagine human-machine autonomy.« *Proceedings of the 2nd Conference on Conversational User Interfaces*, 1–3.

Schlosser, Markus. 2019. »Agency.« In *The Stanford Encyclopedia of Philosophy* (Winter 2019 Edition), edited by Edward N. Zalta. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2019/entries/agency/.

Stark, Birgit, Daniel Stegmann, Melanie Magin, Pascal Jürgens. 2020. »Are algorithms a threat to democracy? The rise of intermediaries: A challenge for public discourse.« *Governing Platforms. Algorithm Watch*. Accessed April 23, 2023. https://algorithmwatch.org/en/wp-content/uploads/2020/05/Governing-Platforms-communications-study-Stark-May-2020-AlgorithmWatch. pdf.

Thimm, Caja, Thomas Christian Bächle. 2019. »Autonomie der Technologie und autonome Systeme als ethische Herausforderung«. In *Maschinenethik: Normative Grenzen autonomer Systeme*, edited by Matthias Rath, Friedrich Krotz, Matthias Karmasin. Springer VS, 73–87.

Verständig, Dan, Christina Kast, Janne Stricker, Andreas Nürnberger (eds.). 2022. *Algorithmen und Autonomie: Interdisziplinäre Perspektiven auf das Verhältnis von Selbstbestimmung und Datenpraktiken*. Opladen: Barbara Budrich.

Weber, Max. 2006 [1948]. *Wirtschaft und Gesellschaft*. Paderborn: Voltmedia.

Weithman, Paul. 2011. »Convergence and political autonomy.« *Public Affairs Quarterly* 25 (4): 327–348.

Whyte, Christopher. 2020. »Deepfake news: AI-enabled disinformation as a multi-level public policy challenge.« *Journal of cyber policy* 5 (2): 199–217.

Witzleb, Norman, Moira Paterson. 2020. »Micro-targeting in Political Campaigns: Political Promise and Democratic Risk.« In *Data-Driven Personalisation in Markets, Politics and Law*, edited by Uta Kohl, Jacob Eisler. Cambridge, 223–240.

Woolley, Samuel C., Philip N. Howard. 2016. »Political communication, computational propaganda, and autonomous agents: Introduction.« *International journal of Communication* 10: 4882–4890.

Zarouali, Brahim, Tom Dobber, Guy De Pauw, Claes de Vreese. 2022. »Using a personality-profiling algorithm to investigate political microtargeting: assessing the persuasion effects of personality-tailored ads on social media.« *Communication Research* 49 (8): 1066–1091.

Autor*innenverzeichnis

Jennifer Burghardt (M.A. Soziale Arbeit) ist zertifizierte Kinderschutzfachkraft und verfügt über langjährige Erfahrung in der Kinder- und Jugendhilfe. In ihrer aktuellen wissenschaftlichen Tätigkeit am Institut für E-Beratung der Technischen Hochschule Nürnberg Georg Simon Ohm erforscht und entwickelt sie gemeinsam mit Partner*innen der Sozialen Arbeit innovative Ansätze für den gemeinwohlorientierten Einsatz von Künstlicher Intelligenz im Kinderschutz sowie in der digitalen psychosozialen Beratung.
E-Mail: jennifer.burghardt@th-nuernberg.de

Cecilia Colloseus (Dr. phil.) ist promovierte Kulturanthropologin. Seit 2022 ist sie wissenschaftliche Mitarbeiterin in der Forschungsgruppe „Human in Command" unter der Leitung von Prof. Doris Aschenbrenner an der Hochschule Aalen. Gegenstand ihrer Forschung ist die gemeinschaftliche und partizipative Entwicklung und Erprobung von KI-Systemen in der „Arbeitswelt der Zukunft".
E-Mail: cecilia.colloseus@hs-aalen.de

David Benjamin Ehrlich (M.Sc.) arbeitete von 2019 bis 2022 als Wissenschaftlicher Mitarbeiter im ökonomischen Teilprojekt des vom BMBF geförderten Forschungsprojekt *CwiC* (*Coping with Certainty*). In dieser Position untersuchte er aus wirtschaftsökonomischer Perspektive, wie Personen mit Ungewissheiten umgehen und welche Kosten sie zu tragen bereit wären, um zusätzliche Informationen zu erhalten oder auch nicht zu erhalten.
E-Mail: david.b.ehrlich@gmail.com

Marc Hauer (M. Sc.) ist Senior Solution Architect im TÜV AI.Lab und beschäftigt sich dort mit den Herausforderungen der Zertifizierung KI-basierter Pro-

dukte. Zusätzlich berät er im Auftrag der Trusted AI GmbH zu Grundlagen und aktuellen Entwicklungen rund um KI. Als medienpädagogischer Referent des Landesmedienzentrums Baden-Württemberg bildet er Schüler, Eltern, Lehrer und Senioren zum Themenkomplex KI an der Schnittstelle zur Gesellschaft weiter. Darüber hinaus engagiert er sich in der Erstellung und Harmonisierung von KI-Normen bei DIN, ETSI und CEN-CENELEC. Seine Dissertation zur Frage wie man Softwareentwicklungsprozesse und primär KI-basierte Softwaresysteme verantwortungsvoll gestalten kann reichte er im Oktober 2023 an der RPTU Kaiserslautern Landau ein.
E-Mail: marc@tuev-lab.ai

Tanja Henking (Prof. Dr. iur., LL.M.) ist seit dem Wintersemester 2015/2016 Professorin für Gesundheitsrecht, Medizinrecht und Strafrecht an der Technischen Hochschule Würzburg-Schweinfurt und leitet dort zudem seit 2019 das im selben Jahr gegründete Institut für Angewandte Sozialwissenschaften (IFAS). Sie forscht zu medizinrechtlichen und -ethischen Fragen am Lebensanfang und am Lebensende, zu Einwilligungsfähigkeit, Zwang und Patientenrechten von Menschen mit psychischer Erkrankung, Digitalisierung und Einsatz künstlicher Intelligenz in der Medizin.
E-Mail: tanja.henking@thws.de

Andreas Hotho (Prof. Dr.) ist Lehrstuhlinhaber für Data Science an der Julius-Maximilians-Universität Würzburg und Sprecher des Center for Artificial Intelligence and Data Science (CAIDAS) der Universität Würzburg. Mit seinem Lehrstuhl forscht er in den letzten Jahren zu datenwissenschaftlichen Themen. Die Kernforschungsbereiche umfassen die Integration von Wissen in Sprachmodelle, die Untersuchung historischer Romane in Kooperation mit den Digital Humanities, die Analyse von Unternehmensdaten für Empfehlungssysteme oder zur Erkennung von Anomalien sowie Untersuchungen basierend auf Sensordaten zur Luftverschmutzung, zur Klimamodellierung und zum Verhalten von Bienen.
E-Mail: hotho@informatik.uni-wuerzburg.de

Sophie Jörg (M.A.) legte ihren Forschungsschwerpunkt nach dem Studium der Politikwissenschaft und Philosophie an der Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU) sowie der Duke University (NC, USA) auf normative Demokratietheorien und politiktheoretische Technikreflexion. Sie lehrte mitunter an der FAU und war als wissenschaftliche Mitarbeiterin in einem

KI-zentrierten Forschungsprojekt an der Hochschule für Philosophie (HFPH) in München tätig. Derzeit arbeitet sie im Planungsstab der Präsidentin des Bayerischen Landtags als Referentin für Strategie und politische Grundsatzfragen. Ihre Dissertation zur Rolle von Plattformdesign im Kontext demokratischen Handelns im digitalen Raum reichte sie im Januar 2024 an der HFPH ein.
E-Mail: joerg.sophie@web.de

Christopher Koska (Dr. phil.) ist wissenschaftlicher Mitarbeiter am Lehrstuhl für Praktische Philosophie der Hochschule für Philosophie München und Partner bei der Unternehmensberatung dimension2 GmbH. Er ist Projektkoordinator des vom bidt finanzierten Forschungsprojekts KAIMo (Kann ein Algorithmus im Konflikt moralisch kalkulieren?) und Postdoc am Center for Responsible AI Technologies (CReAITech). Sein Forschungs- und Arbeitsschwerpunkt ist das Themenfeld der Daten- und Algorithmenethik sowie deren Umsetzung im Kontext der Corporate Digital Responsibility (CDR).
E-Mail: christopher.koska@hfph.de

Maximilian Kraus (M. Sc.) ist wissenschaftlicher Mitarbeiter am Institut für Sozioinformatik der Technischen Hochschule Würzburg-Schweinfurt. Schwerpunkte seiner Forschung sind Künstliche Intelligenz, Mensch-Maschine Interaktionen und Brain-Computer Interfaces. Im Projekt KAIMo untersucht er Möglichkeiten des Einsatzes von KI im Kinderschutz.
E-Mail: maximilian.kraus@thws.de

Nicole Kunkel (Dipl.-theol.) hat in Leipzig, Berlin und Jerusalem studiert und ist derzeit Wissenschaftliche Mitarbeiterin am Lehrstuhl für Ethik und Hermeneutik an der Theologischen Fakultät der Humboldt-Universität zu Berlin. Ihre 2023 verteidigte Dissertation befindet sich derzeit im Veröffentlichungsprozess. In ihr beschäftigt sie sich an der Schnittstelle von Technik- und Friedensethik mit einer ethischen Bewertung autoregulativer Waffensysteme aus bedingt pazifistischer Perspektive.
E-Mail: nicole.kunkel.1@hu-berlin.de

Robert Lehmann (Prof. Dr.) ist Professor für Theorien und Handlungslehre der Sozialen Arbeit an der Technischen Hochschule Nürnberg Georg Simon Ohm und Sprecher der akademischen Leitung des Instituts für E-Beratung. Der Schwerpunkt seiner Lehr- und Forschungstätigkeit liegt bei der psycho-

sozialen Onlineberatung und dem Einsatz von Künstlicher Intelligenz in der Sozialen Arbeit. Näheres unter www.e-beratungsinstitut.de
E-Mail: robert.lehmann@th-nuernberg.de

Nicholas Müller (Prof. Dr.) ist Inhaber der Forschungsprofessur Sozioinformatik und gesellschaftliche Aspekte der Digitalisierung an der Technischen Hochschule Würzburg-Schweinfurt und Leiter des Instituts für Design und Informationssysteme IDIS. Die Professur ist fakultätsübergreifend und das Arbeitsgebiet sowohl in der Informatik und Wirtschaftsinformatik angesiedelt als auch in Forschung und Lehre der Fakultäten Angewandte Sozialwissenschaften und Gestaltung eingebunden. Näheres unter: https://fiw.thws.de/fakultaet/personen/person/prof-dr-habil-nicholas-mueller
E-Mail: nicholas.mueller@thws.de

Carsten Orwat (Dr.) ist wissenschaftlicher Mitarbeiter (senior researcher) am Institut für Technikfolgenabschätzung und Systemanalyse (ITAS) des Karlsruher Instituts für Technologie (KIT). Seit 2000 ist er in zahlreichen Projekten der Technikfolgenabschätzung von Informations- und Kommunikationstechnologien tätig. Weitere Forschungsschwerpunkte sind Governance und Regulierung von Technologien. Derzeit arbeitet er zu den gesellschaftlichen Folgen der Künstlichen Intelligenz, algorithmischen Diskriminierungen und systemischen Risiken.
E-Mail: orwat@kit.edu

Christina Lauppert von Peharnik (Ass. iur.) ist als Vertragsjuristin am Koordinierungszentrum für klinische Studien der Philipps-Universität Marburg tätig. In ihrer vorherigen Position als Wissenschaftliche Mitarbeiterin im BMBF-geförderten Kooperationsprojekt CwiC (Coping with Certainty) von 2019 bis 2022 untersuchte sie die Auswirkungen epistemischer Verschiebungen durch den Einsatz von Künstlicher Intelligenz im Gesundheitssystem aus juristischer Perspektive.
E-Mail: christinaleib@gmx.de

Michael Reder (Prof. Dr. phil.) ist Professor für Praktische Philosophie und Vizepräsident für Forschung an der Hochschule für Philosophie München. Er ist Konsortialführer des vom bidt finanzierten Forschungsverbundes KAIMo (Kann ein Algorithmus im Konflikt moralisch kalkulieren?) und Mitglied des Direktoriums des gemeinsamen Zentrums für verantwortliche KI (CReAI-

Tech) der Technischen Universität München, der Universität Augsburg und der Hochschule für Philosophie München.
E-Mail: michael.reder@hfph.de

Rudolf Seising (PD Dr.) hat Mathematik, Physik und Philosophie an der Ruhr-Universität Bochum studiert und an der Ludwig-Maximilians-Universität (LMU) in München in Wissenschaftstheorie promoviert und in Geschichte der Naturwissenschaften habilitiert. Seit vielen Jahren forscht und lehrt er zur Geschichte der Informatik, der Statistik und der Künstlichen Intelligenz. Nach einigen Auslandsaufenthalten (Österreich, Spanien) und Professurvertretungen an der LMU und an der Friedrich-Schiller-Universität (FSU) in Jena leitete er im Forschungsinstitut des Deutschen Museums von 2019 bis 2023 das wissenschaftshistorische BMBF-Forschungsprojekt „IGGI – Ingenieur-Geist und Geistes-Ingenieure: Eine Geschichte der Künstlichen Intelligenz in der Bundesrepublik Deutschland".
E-Mail: r.seising@deutsches-museum.de

Kerstin Schlögl-Flierl (Prof. Dr. theol.), hat den Lehrstuhl für Moraltheologie an der Universität Augsburg inne. Sie ist seit 2020 Mitglied im Deutschen Ethikrat. Im Bereich der KI-Forschung ist sie eine der Verantwortlichen des Center for Responsible AI Technologies (CReAITech) zwischen Universität Augsburg, Hochschule für Philosophie Münschen und Technische Universität München. Seit 2023 ist sie ein korrespondierendes Mitglied an der Päpstlichen Akademie für das Leben.
E-Mail: kerstin.schloegl-flierl@uni-a.de

Daniel Schlör (Dr. rer. nat.) ist wissenschaftlicher Mitarbeiter am Lehrstuhl für Data Science an der Julius-Maximilians-Universität Würzburg. Seine Lehr- und Forschungsschwerpunkte liegen im Bereich Maschinelles Lernen für Cyber-Security und Anomalie Erkennung. Weitere Forschungsinteressen umfassen auch den Einsatz von Empfehlungssystemen und Natural Language Processing im Bereich medizinischer Informatik.
E-Mail: daniel.schloer@informatik.uni-wuerzburg.de

Jan Fiete Schütte (B.Sc.) ist Berater bei der dimension2 economics & philosophy consult GmbH und verbessert mit Unternehmen datenbasierte Produkte durch die praxisnahe Operationalisierung von Datenethik. Der studierte Sozioinformatiker wirkt außerdem aktiv an der Erstellung und Weiterentwick-

lung von Werkzeugen und Normung mit, wie den CDR Building Bloxx und der DIN NRM II KI. Mit seinem technischen Hintergrund schlägt er die Brücke zwischen Theorie und Praxis, dabei beschäftigt ihn insbesondere die Analyse algorithmischer und selbstlernender Systeme in ihrer Auswirkung auf Mensch und Gesellschaft.
E-Mail: jan-fiete.schuette@dimension2.de

Max Tretter (Mag. theol.) ist Doktorand und wissenschaftlicher Mitarbeiter am Lehrstuhl für Systematische Theologie (Ethik) der Friedrich-Alexander-Universität Erlangen-Nürnberg. Seine Forschungsschwerpunkte umfassen Ethik und Theologie Künstlicher Intelligenz und Robotik, Bioethik des Menschen sowie *Hip Hop Studies*. Weitere Informationen finden Sie online unter: https://www.ethik.phil.fau.de/tretter/
E-Mail: max.tretter@fau.de

Ulrich Freiherr von Ulmenstein (Dipl. iur.) ist zurzeit Rechtsreferendar im Freistaat Sachsen und am Sozialgericht in Leipzig in der 20. Kammer eingesetzt, welche insbesondere Verfahren der Kranken- und Pflegeversicherung bearbeitet. Schwerpunkt seiner bisherigen Forschung war neben datenschutzrechtlichen Grundlagen das Recht der gesetzlichen Krankenversicherung mit seinen verfassungsrechtlichen Grundlagen und einfachgesetzlichen Ausformungen.
E-Mail: u_ulmenstein@outlook.de

Susanna Wolf (M.A.), ist Digitalethikerin und als Data Steward im Data Office bei DATEV eG tätig. Dort berät sie zu wertorientierter Technologiegestaltung und stärkt gemeinsam mit ihren Kolleg:innen Good Data Governance: Neben Effizienz und Wertschöpfung ist die Wertperspektive im Sinne Digitaler Verantwortung (Corporate Digital Responsibility, kurz: CDR) hier ebenso im Fokus wie partnerschaftliche Vernetzung mit Bezug auf vertrauenswürdige Datenräume insbesondere im europäischen Kontext. Als Teil von Good Data Governance steht auch Datenethik@DATEV für gewinnbringende Vernetzung und wertorientierte Innovation. Susanna Wolf hat das Thema von der partizipativen Entwicklung hin zur internen Verankerung fachlich geführt. Für Datenethik@DATEV hat die Genossenschaft 2021 den CDR-Award erhalten und unterstützt die CDR-Community als Good Practice. Neben ihrer Tätigkeit bei DATEV promoviert Susanna Wolf derzeit zur Bedeutung von CDR-Regulierungsbestrebungen und deren Umsetzbarkeit für die Zukunft von vertrauens-

würdiger KI.
E-Mail: susanna.wolf@datev.de

Paula Ziethmann (M.A.) ist wissenschaftliche Mitarbeiterin am Center for Responsible AI Technologies (CReAITech) und Doktorandin der Technikphilosophie an der Universität Augsburg. In Ihrer Dissertation untersucht sie den Einsatz von Künstlicher Intelligenz in der Medizin. Dabei verbindet sie Theorien von Michel Foucault mit qualitativer Sozialforschung und technikphilosophischen Überlegungen.
E-Mail: paula.ziethmann@zig.uni-augsburg.de