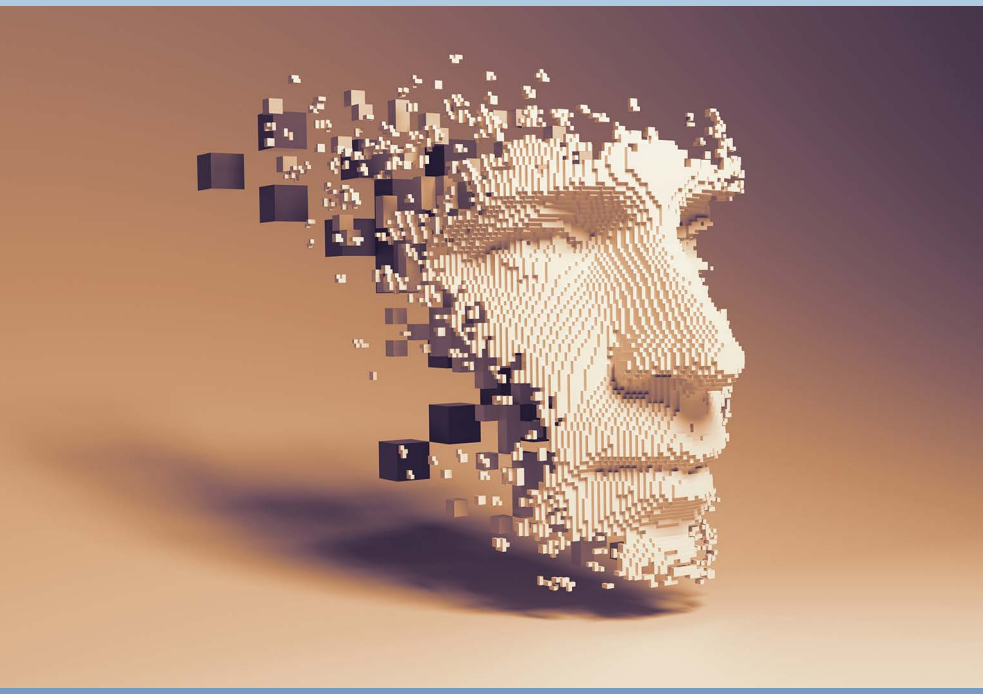
Deutscher Ethikrat



Mensch und Maschine – Herausforderungen durch Künstliche Intelligenz

STELLUNGNAHME

Deutscher Ethikrat

Mensch und Maschine – Herausforderungen durch Künstliche Intelligenz

STELLUNGNAHME

20. März 2023

Herausgegeben vom Deutschen Ethikrat

Jägerstraße 22/23 · D-10117 Berlin
Telefon: +49/30/20370-242 · Telefax: +49/30/20370-252
E-Mail: kontakt@ethikrat.org
www.ethikrat.org

2. Auflage, 2025
© 2023 Deutscher Ethikrat, Berlin
Alle Rechte vorbehalten.
Eine Abdruckgenehmigung wird auf Anfrage gern erteilt.
Layout: Torsten Kulick
Titelillustration: pinkeyes/Shutterstock.com

DANKSAGUNG

An dieser Stellungnahme haben zahlreiche Personen unterstützend mitgewirkt:

Ulrike von Luxburg, Matthias Bethge, Tanja Schultz und Stefan Remy boten zum Projektbeginn bei einer öffentlichen Anhörung am 25. Februar 2021 wertvolle Orientierung zu den aktuellen und absehbaren Entwicklungen im Bereich Künstlicher Intelligenz. Inge Molenaar und Ute Schmid ergänzten dies im Rahmen einer weiteren, nicht öffentlichen Anhörung sowie mit schriftlicher Expertise zum Einsatz von KI in der Schule und Matthias May gewährte hilfreiche Einblicke in die Anwendung algorithmischer Systeme in der medizinischen Diagnostik.

Wolfgang Schulz, Indra Spiecker genannt Döhmann und noch einmal Ute Schmid brachten ihren Sachverstand zu juristischen und informatikspezifischen Passagen der Stellungnahme ein. Carl Friedrich Gethmann begleitete die Arbeit der Stellungnahme auch nach dem Ende seiner Amtszeit im Deutschen Ethikrat im Februar 2021 als nunmehr externes Mitglied der Arbeitsgruppe „Mensch und Maschine“ und auch Andreas Kruse wirkte bis zu seinem Ausscheiden aus dem Ethikrat im März 2022 an der Stellungnahme mit.

Der Deutsche Ethikrat dankt allen Mitwirkenden sehr herzlich für ihre Unterstützung!

INHALTSVERZEICHNIS

ZUSAMMENFASSUNG 11

1 EINLEITUNG 78

TEIL I: TECHNISCHE UND PHILOSOPHISCHE GRUNDLEGUNGEN

2 ZENTRALE ENTWICKLUNGEN UND TECHNISCHE GRUNDLAGEN KÜNSTLICHER INTELLIGENZ 86

2.1 Historischer Kontext 86

2.2 KI im 21. Jahrhundert: Big Data, Algorithmen und soziotechnische Ökosysteme 92

2.2.1 Digitale Durchdringung der menschlichen Lebenswelt 92

2.2.2 Daten und digitale Infrastrukturen 94

2.2.3 Algorithmen und Datenverarbeitung 97

2.2.4 Einsatzbereiche algorithmischer Systeme und KI 105

2.2.5 Ethische Leitlinien und regulativer Rahmen für algorithmische Systeme und KI 108

3 ZENTRALE BEGRIFFE UND PHILOSOPHISCHE GRUNDLAGEN 116

3.1 Künstliche Intelligenz: begriffliche Analyse 116

3.2 Intelligenz und Vernunft 123

3.2.1 Intelligenz 123

3.2.2 Vernunft 129

3.3 Handlung und Verantwortung 135

3.3.1 Handlung 135

3.3.2 Verantwortung 140

3.4 Anthropologische Aspekte des Mensch-Maschine-Verhältnisses 145

3.4.1 Philosophische Grundbestimmung des Menschseins 145

3.4.2 Der Mensch als Maschine – die Maschine als Mensch? 149

3.4.3 Verleiblichte Vernunft 156

3.5 Fazit 161

4 MENSCH-TECHNIK-RELATIONEN 164

4.1 Einleitung 164

4.2 Technikdeterminismus versus Sozialkonstruktivismus 165

4.3 Mehrstufige Mensch-Technik-Wechselwirkungen 170
4.4 Erweiterung und Verminderung menschlicher Autonomie und Autorschaft 177
4.5 Fazit 183

TEIL II: AUSGEWÄHLTE ANWENDUNGEN UND SEKTORSPEZIFISCHE EMPFEHLUNGEN

5 MEDIZIN 190
5.1 Einleitung 190
5.2 Einsatz von KI-Systemen in der Medizin 192
5.2.1 Entwicklung von KI-Systemen 192
5.2.2 KI in der medizinischen Forschung 196
5.2.3 KI in der medizinischen Versorgung 199
5.3 Fazit und Empfehlungen 215

6 BILDUNG 219
6.1 Einleitung 219
6.2 Zum Bildungsbegriff 221
6.3 Einsatzmöglichkeiten datenbasierter, KI-gestützter Lehr- und Lernsysteme 225
6.4 Mensch-Maschine-Relationen in der schulischen Bildung: Erweitern, Vermindern, Ersetzen 228
6.5 Grundsätzliche Diskussion: KI im schulischen Bildungsprozess 239
6.6 Fazit und Empfehlungen 244

7 ÖFFENTLICHE KOMMUNIKATION UND MEINUNGSBILDUNG 251
7.1 Einleitung 251
7.2 Das Internet als soziotechnisches System: Funktionsweise sozialer Medien 252
7.2.1 Neue soziotechnische Infrastrukturen 252
7.2.2 Informationsauswahl und Kuratierung 254
7.2.3 Moderation von Inhalten 258
7.2.4 Auswirkungen auf die Erweiterung und Verminderung menschlicher Handlungsfähigkeiten 262
7.3 Informationsqualität 264
7.3.1 Falschnachrichten und Verschwörungstheorien 266
7.3.2 Filterblasen und Echokammern 268

7.3.3 Moralische und emotionale Aufladung 270
7.3.4 Relevanz der beobachteten Effekte 272
7.4 Diskursqualität 273
7.4.1 Politische Polarisierung 275
7.4.2 Politische Werbung und Manipulation 279
7.4.3 Spannungsfeld Diskursverrohung und Äußerungsfreiheit 284
7.4.4 Erweiternde und vermindernde Rückwirkungen auf den öffentlichen Vernunftgebrauch 288
7.5 Fazit und Empfehlungen 289
8 ÖFFENTLICHE VERWALTUNG 298
8.1 Einleitung 298
8.2 Ethische Fragen algorithmischer Automatisierung im Verwaltungshandeln 299
8.3 Automatische Entscheidungssysteme am Beispiel des Sozialwesens 306
8.3.1 Erweiterung professioneller Handlungskompetenz 307
8.3.2 Verminderung von Entscheidungskompetenz, Handlungsoptionen und Autorschaft 313
8.4 Predictive Policing – KI in der Kriminalitätsbekämpfung 320
8.5 Fazit und Empfehlungen 327

TEIL III: QUERSCHNITTSTHEMEN UND ÜBERGREIFENDE EMPFEHLUNGEN

9 ZUSAMMENFASSUNG DER BISHERIGEN ANALYSE 334
9.1 Anthropologische und ethische Grundorientierung 334
9.2 Einsichten aus den Anwendungsfeldern 339
10 ENTFALTUNG VON QUERSCHNITTSTHEMEN UND EMPFEHLUNGEN 345
10.1 Querschnittsthema 1: Erweiterung und Verminderung von Handlungsmöglichkeiten 346
10.2 Querschnittsthema 2: Wissenserzeugung durch KI und Umgang mit KI-gestützten Voraussagen 348
10.3 Querschnittsthema 3: Gefährdung des Individuums durch statistische Stratifizierung 351
10.4 Querschnittsthema 4: Auswirkungen von KI auf menschliche Kompetenzen und Fertigkeiten 353

10.5 Querschnittsthema 5: Schutz von Privatsphäre und Autonomie versus Gefahren durch Überwachung und Chilling-Effekte 357
10.6 Querschnittsthema 6: Datensouveränität und gemeinwohlorientierte Datennutzung 361
10.7 Querschnittsthema 7: Kritische Infrastrukturen, Abhängigkeiten und Resilienz 366
10.8 Querschnittsthema 8: Pfadabhängigkeiten, Zweitverwertung und Missbrauchsgefahren 369
10.9 Querschnittsthema 9: Bias und Diskriminierung 372
10.10 Querschnittsthema 10: Transparenz und Nachvollziehbarkeit – Kontrolle und Verantwortung 376
10.11 Fazit 379

LITERATURVERZEICHNIS 381

ABKÜRZUNGSVERZEICHNIS 402

ZUSAMMENFASSUNG

Einleitung

1) Digitale Technologien und Künstliche Intelligenz (KI) haben mittlerweile in nahezu allen Bereichen des öffentlichen und privaten Lebens Einzug gehalten. Für die ethische Bewertung solcher Entwicklungen und ihres Einsatzes in verschiedenen Bereichen ist es nötig, nicht nur die Technologien zu verstehen, sondern auch ihre Wechselwirkungen mit den Personen, die sie verwenden oder von ihrer Anwendung betroffen sind. Zentral ist dabei die Frage, welche Auswirkungen damit verbunden sind, wenn Tätigkeiten, welche zuvor Menschen vorbehalten waren, an Maschinen delegiert werden. Werden menschliche Autorschaft und Handlungsmöglichkeiten durch den Einsatz von KI erweitert oder vermindert?
2) In der vorliegenden Stellungnahme geht der Deutsche Ethikrat dieser Frage nach und schreibt damit Themen fort, die bereits in den Stellungnahmen „Big Data und Gesundheit – Datensouveränität als informationelle Freiheitsgestaltung“ (2017) sowie „Robotik für gute Pflege“ (2020) angeschnitten wurden. Er reagiert mit den aktuellen Ausführungen zudem auf eine im Oktober 2020 vom Präsidenten des Deutschen Bundestages formulierte Bitte, eine multidisziplinäre Stellungnahme zu den ethischen Fragen des Verhältnisses von Mensch und Maschine zu erarbeiten.
3) Die Stellungnahme gliedert sich in drei Teile. Im ersten Teil geht es um die *technischen und philosophischen Grundlagen* des Themas. Im zweiten Teil werden die zuvor angestellten Überlegungen anhand von ethischen Analysen in vier *ausgewählten Anwendungsfeldern* exemplarisch konkretisiert: der Medizin, der schulischen Bildung, der öffentlichen Kommunikation und Meinungsbildung sowie der öffentlichen Verwaltung. Im dritten

Teil werden zehn in allen Anwendungsbereichen relevante *Querschnittsthemen* entfaltet, welche jeweils auch übergreifende Empfehlungen enthalten.

TEIL I: TECHNISCHE UND PHILOSOPHISCHE GRUNDLEGUNGEN

Zentrale Entwicklungen und technische Grundlagen Künstlicher Intelligenz

4) Die Idee von Maschinen, deren Fähigkeiten in bestimmten, für das menschliche Wesen besonders prägenden Kernbereichen wie dem Erkennen, Lernen oder Handeln menschlichen Fähigkeiten ähneln oder diese sogar übertreffen, lässt sich bis in die griechische Mythologie zurückverfolgen, Jahrtausende vor der Erfindung von Softwaresystemen. Mit dem Bau der ersten Computer im 20. Jahrhundert rückte die Existenz maschineller Intelligenz erstmals in greifbare Nähe. Der Mathematiker Alan Turing formulierte 1950 ein später als Turing-Test bezeichnetes Kriterium für KI, nach dem maschinelle Intelligenz dann vorläge, wenn das Verhalten einer Maschine für menschliche Beobachter nicht von dem eines Menschen unterscheidbar erscheint.

5) Die frühe Forschung zu KI ging davon aus, dass man menschliches Lernen oder menschliche Intelligenz so genau beschreiben könne, dass eine Maschine dazu gebracht werden kann, sie zu simulieren. Konkrete KI-Forschungsthemen, die bis heute eine Rolle spielen, waren von Anfang an zum Beispiel Mustererkennung, Sprachverarbeitung, Abstraktionsfähigkeit, Kreativität und flexibles Problemlösen. Flankiert von Fortschritten bei der Entwicklung von Computerhardware und Programmiersprachen entstand bald großer Optimismus über die Potenziale maschineller Intelligenz. In den folgenden Jahrzehnten wechselten sich enthusiastische Phasen mit

sogenannten „KI-Wintern“ ab, während derer Enttäuschungen über vermeintlich ausbleibende praktische Erfolge im Vordergrund standen und Fördermittel gekürzt wurden.

6) Weiterentwicklungen in einzelnen KI-Kerngebieten, die Entstehung paralleler Datenverarbeitungsmethoden und des Internets sowie das wachsende Engagement von Forschungsorganisationen, Militär und Industrie prägten die KI-Forschung im späten 20. Jahrhundert. Parallel dazu entstand ein kritischer Diskurs, im Zuge dessen sich auch die Computerethik als eigene Disziplin etablierte. Dabei kamen zunehmend auch philosophische Zweifel auf, ob insbesondere die von einigen Forschenden in Aussicht gestellten Visionen einer generellen oder starken KI jemals realisiert werden könnten – oder sollten.

7) Um die Jahrtausendwende nahmen drei Entwicklungen Fahrt auf, die der Entwicklung von KI zu einer bis heute anhaltenden Dynamik verhalfen: erstens eine deutliche Leistungssteigerung und Miniaturisierung von Computern, zweitens eine zunehmende Vernetzung digitaler Systeme und drittens damit verbundene neue Möglichkeiten der Datenzusammenführung und -auswertung.

8) Diese Entwicklungen haben zu einer intensivierten Durchdringung der Alltagswelt mit Computern geführt, darunter auch zahlreiche vernetzte und mit Sensoren versehene „smarte“ Alltagsbegleiter wie Mobiltelefone, Uhren und Haushaltsgeräte. Es entstehen soziotechnische Datenökosysteme, in denen über Geräte und die sie verknüpfenden Datennetzwerke zunehmend akkurate und umfangreiche digitale Repräsentationen der Bewegungen, Handlungen, Eigenschaften und Präferenzen vieler Personen entstehen. Solche digitalen Abbilder können nicht nur ausgewertet werden, sondern wirken auch auf menschliches Verhalten zurück, indem auf ihrer Grundlage Menschen Informationen oder Handlungsempfehlungen angeboten werden.

9) Das Fundament solcher digitalen Operationen und Interaktionen bilden Daten und die sie begleitenden Metadaten, die jeweils von höchst unterschiedlicher Natur wie Qualität sein können. Die Qualität eines Datensatzes hängt dabei nicht nur davon ab, wie genau, vollständig, aktuell oder detailliert seine Daten sind, sondern auch vom Verhältnis zwischen den Erhebungs- und den Anwendungskontexten. Daten können einer jeweiligen Frage oder Aufgabenstellung mehr oder weniger angemessen sein. Werden solche Fragen von Qualität und Passung nicht rechtzeitig und hinreichend berücksichtigt, sind Fehler, Verzerrungen (Bias) und irreführende Analysen möglich.

10) Entscheidend für die Leistungsfähigkeit datengetriebener Anwendungen sind auch die Hardware und Infrastruktur, die für die Handhabung und Nutzung von Daten zur Verfügung stehen. Hier kommen aktuell sowohl über das Internet zugängliche Dienste zum Einsatz (Cloud-Computing), hinter denen Großeinrichtungen stehen, die auf Datenspeicherung und/oder Datenanalyse spezialisiert sind, wie auch immer leistungsfähigere Möglichkeiten, Daten zumindest teilweise bereits lokal in den Geräten, die sie erheben, zu verarbeiten (Edge-Computing).

11) Herzstück jeglicher Datenverarbeitung sind Algorithmen: Verarbeitungsanweisungen, die vorgeben, wie eingegebene Daten meist schrittweise nach klar definierten Regeln umgeformt werden, bis der gesuchte Ausgabewert erreicht ist. Im Kontext aktueller KI-Forschung sind statistische Analysen, mit denen Regelmäßigkeiten in Daten erkannt sowie Zusammenhänge zwischen einzelnen Merkmalen identifiziert werden, von besonderer Bedeutung. Auf dieser Grundlage können Vorhersagen für ähnliche Datensätze oder künftige Entwicklungen abgeleitet werden. Geht es darum, kausale Mechanismen nachzuweisen, sind in der Regel weitere Überlegungen und Untersuchungen nötig, die eine plausible Erklärung

für den vermuteten Wirkzusammenhang zwischen Merkmalen anbieten, die sich auch empirisch – zum Beispiel in Experimenten – überprüfen lässt.

12) Statistische Analysen enthalten Unsicherheiten, die sich in der Regel nicht ganz ausmerzen lassen. Mit der Minimierung bestimmter Fehlerquellen können zudem andere Fehlerquellen verstärkt werden. Welche Fehler in statistischen Analysen am ehesten in Kauf zu nehmen sind, hängt daher auch immer von der konkreten Fragestellung und Zwecksetzung ab und ist in zahlreichen Bereichen nicht nur eine technisch-methodische, sondern auch eine ethische Frage.

13) Die in KI-Systemen verwendeten algorithmischen Verfahren und Systeme werden vielfach unter dem Stichwort „maschinelles Lernen" zusammengefasst und zeichnen sich dadurch aus, dass sie ihre Mustersuche, Modellbildung und sonstige Funktionsweise datenbasiert optimieren können. Dabei gibt es anfangs eine Trainingsphase, in der ein Algorithmus sein Modell zur Mustererkennung durch wiederholte Analyse von Trainingsdaten aufbaut und verfeinert.

14) Maschinelles Lernen umfasst unterschiedliche Ansätze. Beim überwachten Lernen sind die Zuordnungen zwischen den Eingabe- und den gesuchten Ausgabedaten im Trainingsdatensatz bereits bekannt, beispielsweise Bilder von gesunder Haut und Hautkrebs, deren gesicherte Zuordnung zu einer dieser beiden Kategorien in einem Etikett (Label) vermerkt ist. Unüberwachtes Lernen hingegen funktioniert ohne vorherige Etikettierung der Trainingsdaten; stattdessen „sucht" der Algorithmus eigenständig nach Mustern in Daten. Beim Verstärkungslernen optimiert der Algorithmus seine Operationen auf bestimmte Ziele hin und erhält dabei in der Trainingsphase für jeden Versuch eine Rückmeldung, ob dieser Schritt das System dem Ziel nähergebracht oder es davon entfernt hat.

15) Deep Learning ist ein Teilbereich des maschinellen Lernens, der besonders für den Umgang mit großen Datenmengen geeignet ist und in den letzten Jahren zu einem wichtigen Treiber für viele KI-Anwendungen geworden ist. Hier kommen sogenannte neuronale Netze zum Einsatz, deren Funktionsweise entfernt an Netzwerkstrukturen im Gehirn angelehnt ist.

16) Die algorithmischen Strategien, die im Laufe des Trainings zur Bewältigung der jeweiligen Aufgaben entwickelt werden, sind in der Regel selbst für geschultes Personal, das den Code einsehen kann, nicht vollständig nachvollziehbar (Blackbox). Es gibt verschiedene Lösungsansätze, um trotzdem eine für die jeweilige Zielgruppe angemessene Transparenz, Interpretierbarkeit oder Erklärbarkeit algorithmischer Prozesse zu erreichen (Explainable AI). Deren Auswahl und Anwendung ist jedoch technisch anspruchsvoll.

17) Die kombinierten Entwicklungen von Hardware und Software, Vernetzung und Datenproduktion haben vielfältig einsetzbare Anwendungsmöglichkeiten von KI hervorgebracht. KI-Systeme können beispielsweise inzwischen Menschen in anspruchsvollen Strategiespielen wie Schach und Go schlagen (MuZero) oder komplexe Texte produzieren, deren maschineller Ursprung oftmals nicht mehr zu erkennen ist (ChatGPT).

18) Der Deutsche Ethikrat nimmt in dieser Stellungnahme vier Handlungsfelder in den Blick, in denen der Einsatz von KI entweder schon besonders weitreichende Veränderungen mit sich bringt oder dies in näherer Zukunft bewirken könnte. In der *Medizin* stellt maschinelles Lernen beispielsweise Fortschritte bei Diagnostik und individualisierten Präventions- und Therapieempfehlungen in Aussicht. In der schulischen *Bildung* entstehen vielfältige Ansätze, die Vermittlung von Wissen und Kompetenzen in der Schule mithilfe von KI effektiver zu gestalten. In der *öffentlichen Kommunikation und Meinungsbildung*

läuft inzwischen ein Großteil des Informationsaustauschs über digitale Plattformen und soziale Medien ab. In der *öffentlichen Verwaltung* berührt der Einsatz algorithmisch gestützter Entscheidungshilfen und Prognosen das Leben vieler Menschen, beispielsweise bei der Beurteilung oder Überwachung von Personen im Bereich des Sozial- oder Polizeiwesens.

19) Um auf die mit solchen Veränderungen verbundenen Herausforderungen für das menschliche Miteinander zu reagieren, sind bereits eine Reihe von Regularien entstanden oder aktuell in der Entwicklung. Dazu gehört zum einen eine Fülle an Leitlinien, die von Kodizes einzelner Unternehmen über Richtlinien von Fachgesellschaften bis hin zu Werken auf nationaler oder internationaler Ebene reicht. Parallel entwickelt sich zum anderen auch der regulative Rahmen weiter, beispielsweise in den Mediengesetzen in Deutschland. Da soziotechnische Entwicklungen im KI-Bereich häufig von international agierenden IT-Unternehmen vorangetrieben werden, gewinnen Regelungen auf übernationaler Ebene zunehmend an Bedeutung, in der Europäischen Union beispielsweise die Datenschutz-Grundverordnung und der Vorschlag für einen Artificial Intelligence Act.

20) Diese Entwicklungen berücksichtigend fokussiert der Deutsche Ethikrat nicht auf den rechtlichen Rahmen, sondern gründet seine Analyse der Konsequenzen digitaler Entwicklungen für das menschliche Zusammenleben auf einer philosophischen Auseinandersetzung mit den anthropologischen Grundbegriffen, die im Mittelpunkt des menschlichen Selbstverständnisses stehen. Darauf aufbauend entwickelt er ein Verständnis von Mensch-Technik-Relationen, in dem es entscheidend darauf ankommt, wie die Delegation menschlicher Tätigkeiten an Maschinen und algorithmische Systeme auf zentrale anthropologische Konzepte zurückwirkt und dabei

insbesondere menschliche Autorschaft erweitert oder vermindert.

Zentrale Begriffe und philosophische Grundlagen

21) Das Verständnis des Begriffs der Künstlichen Intelligenz hat sich im Laufe der Jahre verändert und unterscheidet sich sowohl innerhalb als auch zwischen verschiedenen Berufsgruppen und Disziplinen. Eine große Rolle spielt die Unterscheidung zwischen sogenannter schwacher und starker KI, wobei letztere Vision eine menschenähnliche oder gar menschliche Fähigkeiten übertreffende KI beschreibt. Weitere Begriffspaare, mit denen unterschiedliche Formen oder Grade der Annäherung von KI an menschliche Intelligenz erfasst werden sollen, sind spezielle versus allgemeine KI sowie enge versus breite KI.

22) Die Charakterisierung als spezielle, enge oder schwache KI einerseits sowie allgemeine, breite oder starke KI andererseits verweist nicht nur jeweils auf Differenzen zwischen zwei Polen. Dahinter verstecken sich, insbesondere beim Begriffspaar der schwachen und starken KI, vielmehr auch unterschiedliche Verständnisse von Intelligenz sowie unterschiedliche Positionen hinsichtlich der Kernfrage, ob es qualitative und kategorische oder nur quantitative und prinzipiell überwindbare Unterschiede zwischen menschlicher Intelligenz und KI gibt.

23) Wichtig ist zum einen die Differenz hinsichtlich der Breite bzw. Enge des Fähigkeitenspektrums der KI. Die meisten KI-Anwendungen entfalten ihre jeweilige Leistung auf klar umrissenen, engen Gebieten oder Domänen. Zum anderen geht es jedoch auch darum, ob Intelligenz an bestimmte mentale Voraussetzungen geknüpft ist, welche über die bloße *Simulation* von Verständnis hinausgehen. Es ergibt sich also die Frage, ob Intelligenz in allgemeiner oder starker Form jemals vollumfänglich

Maschinen zukommen kann oder ob dafür spezifisch menschliche Eigenschaften Voraussetzung sind.

24) Antworten auf diese Frage variieren in unterschiedlichen anthropologischen Theoriemodellen. Aus behavioristischer Sicht würden manche in einem humanoiden Roboter mit perfekten Bewegungsfähigkeiten und einer menschenähnlichen Mimik und Gestik ein Beispiel breiter oder gar starker KI sehen, wenn er in der Lage wäre, alle menschlichen kognitiven Fähigkeiten perfekt zu simulieren. Nach anderen Konzeptionen wäre hingegen zu bestreiten, dass damit eine Form starker KI vorliege, da auch eine perfekte Simulation nicht garantiere, dass ein solcher Roboter mentale Zustände aufweise, über Einsichts- und Urteilsfähigkeit sowie über emotive Einstellungen wie Hoffnungen und Ängste verfüge.

25) In dieser Stellungnahme wird vorausgesetzt, dass die Unterscheidung zwischen enger und breiter KI quantitativer bzw. gradueller Natur ist, die Entstehung einer starken KI jedoch einen qualitativen Sprung bedeuten würde. Als *enge KI* gelten dabei Anwendungen, welche menschliche Fähigkeiten in einer Domäne simulieren, um spezifische Aufgaben zu erfüllen. *Breite KI* erweitert das Spektrum ihrer Anwendbarkeit über einzelne Domänen hinaus. Der Begriff der *starken KI* wird für die Vision einer Künstlichen Intelligenz verwendet, die jenseits der möglicherweise perfekten Simulation menschlicher Kognition auch über mentale Zustände, Einsichtsfähigkeit und Emotionen verfügen würde.

26) Eine wichtige Grundlage für Diskussionen über die aktuellen und möglichen künftigen Potenziale von KI sind Vorstellungen zu menschlicher Intelligenz. Aus psychologischer Perspektive ist Intelligenz als ein hypothetisches Konstrukt aufzufassen, das als solches zwar verbal umschrieben werden kann, zum Beispiel im Sinne von Verstehen, Urteilen und Schlussfolgern oder zielgerichtetem Handeln, rationalem Denken und effektiver

Auseinandersetzung mit der Umwelt, aber nicht direkt beobachtbar ist. Intelligenztests ermöglichen hier eine Operationalisierung, indem sie Situationen anbieten, in denen Menschen Verhalten zeigen können, das vor dem Hintergrund eines theoretischen Vorverständnisses als mehr oder weniger „intelligent" bezeichnet werden kann.

27) Die Frage, ob Intelligenz eine einheitliche Fähigkeit ist oder viele Fähigkeiten umfasst, die gegebenenfalls auch voneinander unabhängig sein können, ist empirisch nicht eindeutig zu klären. Viel diskutiert und mit Blick auf KI von Relevanz ist auch der Zusammenhang von Intelligenz und Kreativität. Eine wichtige Rolle spielt hierbei die Unterscheidung zwischen konvergentem Denken, das durch logische Schlussfolgerungen zu einer einzigen oder besten Lösung gelangt, und dem für Kreativität charakteristischen divergenten Denken, das mehrere alternative Lösungen finden kann, die jeweils den gegebenen Anforderungen entsprechen.

28) In jüngerer Zeit hat sich der Blick auf Intelligenz sukzessive erweitert, beispielsweise mit Konzepten wie die der sozialen bzw. emotionalen Intelligenz. Darüber hinaus entwickelte sich rund um die Stichworte *embodied*, *embedded*, *enactive* und *extended* Kognition ein Forschungsfeld, welches in Philosophie, Psychologie und Robotik die Rolle des Körpers einerseits sowie der Umwelt andererseits für Intelligenz und kognitive Leistungen erforscht. Spätestens diese Erweiterungen werfen die grundsätzliche Frage auf, wie die Übertragung des Intelligenzbegriffs auf technische Artefakte zu verstehen ist. Man sollte daher die Verwendung des Ausdrucks „Intelligenz" in der Wortverbindung „Künstliche Intelligenz" eher als eine Metapher einordnen, deren Beschreibungs- und Erklärungsfunktion genauerer Aufklärung bedarf.

29) Der Begriff der Vernunft wurde bereits lange vor der Einführung des Begriffs der Intelligenz verwendet, um die spezifische menschliche Fähigkeit zu kennzeichnen,

sich in der Welt zu orientieren, selbstverantwortlich zu handeln und so der eigenen Lebenspraxis eine kohärente Struktur zu geben. Intelligenz ist für Vernunft eine wichtige Voraussetzung, aber keine hinreichende Bedingung.

30) Der Vernunftbegriff ist überaus komplex und umfasst ein mehrdimensionales Beziehungsgefüge von Denk-, Reflexions- und Operationsformen, das in seiner Gesamtheit im Dienste einer möglichst adäquaten Wirklichkeitserschließung steht und in einen komplexen sozialen und kulturellen Kontext verwoben ist. Von grundlegender Bedeutung ist dabei die Gegenüberstellung von theoretischer Vernunft, die sich auf den Erkenntnisgewinn richtet, um zu wahren empirischen oder apriorischen Urteilen zu gelangen, und praktischer Vernunft, die auf ein kohärentes, verantwortliches Handeln abzielt, um ein gutes Leben zu ermöglichen.

31) Vor allem im Blick auf theoretische Vernunft scheinen sich einige Parallelen zur Arbeitsweise von KI-Systemen aufzudrängen. In beiden Bereichen spielen Fähigkeiten der Informationsverarbeitung, des Lernens, des logischen Schlussfolgerns und konsistenten Regelfolgens sowie der sinnvollen Verknüpfung gespeicherter Informationen eine zentrale Rolle. Bei näherer Betrachtung zeigen sich jedoch insofern gravierende Differenzen, als sich nicht nur die Arbeitsweise des menschlichen Gedächtnisses in mehrfacher Hinsicht vom technischen Speicher eines Computers unterscheidet, sondern auch die menschliche Urteilspraxis technisch nicht substituierbar ist. Zumindest die bislang verfügbaren KI-Systeme verfügen nicht über die dafür relevanten Fähigkeiten des Sinnverstehens, der Intentionalität und der Referenz auf eine außersprachliche Wirklichkeit.

32) Dies bestätigt sich erst recht für die praktische Vernunft, die insofern noch von weit komplexerer Natur ist, als ihr Ziel nicht nur in wohlbegründeten praktischen Einzelurteilen, sondern in einem möglichst richtigen und

verantwortlichen Handeln besteht, das über einen langen Zeitraum aufrechterhalten wird, eine kohärente Ordnung der Praxis garantiert und damit ein insgesamt gutes Leben ermöglicht. Dazu bedarf es mehrerer Einzelkompetenzen, deren Simulationsmöglichkeiten durch technische Artefakte kontrovers diskutiert werden.

33) Zu diesen Einzelkompetenzen gehören erstens ein Verständnis der für unsere Moralsprache bedeutsamen Begriffe zur Bezeichnung moralisch relevanter Güter, Werte und Haltungen, zweitens ein Unterscheidungs- und Einfühlungsvermögen, drittens die Fähigkeit zur Abwägung konfligierender Güter und Werte, viertens die Befähigung zum reflektierten Umgang mit Regeln unterschiedlicher Reichweite, fünftens die Kompetenz zum intuitiven Erfassen komplexer Handlungssituationen und Umstände, sechstens ein Urteilsvermögen, siebtens die Fähigkeit zur Begründung der eigenen moralischen Urteile und der ihnen korrespondierenden Praxis und achtens eine Affekt- und Impulskontrolle, um die jeweils gefällten praktischen Urteile auch handlungswirksam werden zu lassen.

34) Während partielle Überschneidungen des Kompetenzprofils moderner KI-Systeme mit dem komplexen Phänomen menschlicher Vernunft durchaus möglich sind, ist zu berücksichtigen, dass die hier genannten Einzelfähigkeiten nicht beziehungslos nebeneinanderstehen. Vielmehr ist von vielfältigen Wechselwirkungen, Rückkopplungen und Bedingungsverhältnissen zwischen ihnen auszugehen. Sie bilden einen integralen Bestandteil einer komplexen menschlichen Natur, die als leibseelische Einheit zu verstehen ist. Menschliche Vernunft ist stets als verleiblichte Vernunft zu begreifen. Praktische Vernunft ist zudem nicht aus einer rein individualistischen Perspektive zu verstehen. Da jeder Mensch Teil einer sozialen Mitwelt und kulturellen Umgebung ist, die sich nachhaltig auf seine Sozialisation auswirken, müssen

auch überindividuelle kulturelle Faktoren in die Deutung der praktischen Vernunft einbezogen werden.

35) Ein angemessenes Verständnis insbesondere des praktischen Vernunftgebrauchs ist eng mit unserem basalen Selbstverständnis als handlungsfähige Personen verbunden. Nicht jedes menschliche Tun, das auf die Umwelt einwirkt, ist als Handlung zu verstehen, sondern nur solches, das zweckgerichtet, beabsichtigt und kontrolliert ist. Unterstellt man, dass Maschinen nicht zweckgerichtet operieren, also keine Absichten haben, dann ist die Zuschreibung von Handlungen in Bezug auf Maschinen in diesem engen Sinne nicht möglich.

36) Im KI-Diskurs kommt allerdings seit der Jahrtausendwende zunehmend die Frage auf, in welchem Sinne Maschinen außerhalb des obigen engen Handlungsbegriffs doch in bestimmten Kontexten in einem weiteren Sinne handeln können, zum Beispiel wenn Entscheidungen komplett an Softwaresysteme delegiert werden. Daran anknüpfend gibt es einen Diskurs, ob und inwieweit zunehmend eigenständige, das heißt ohne menschliches Zutun funktionierende maschinelle Systeme als „Agenten" in der Folge für ihr „Handeln" verantwortlich gemacht werden können, etwa mit Blick auf Fragen der Haftung.

37) Selbst wenn Maschinen komplexe Vollzüge oder Operationen durchführen, damit Veränderungen in der Welt bewirken und flexibel mit anspruchsvollen Herausforderungen der menschlichen Lebenswelt umgehen können, führen sie diese Veränderungen aber nicht absichtlich herbei und haben sie diese daher auch nicht in einem moralischen und rechtlichen Sinne zu verantworten. Vor diesem Hintergrund scheint es sinnvoll, den Handlungsbegriff im engen Sinne Menschen vorzubehalten, um inflationären Ausweitungen des Akteursstatus zu vermeiden und konzeptionelle Grenzziehungen zu ermöglichen.

38) Entscheidend ist demnach das Konzept der Handlungsurheberschaft bzw. Autorschaft, das auf die universelle menschliche Erfahrung verweist, sich selbst und andere im Hinblick auf bestimmte Ereignisse und Zustände als Urheber anzusehen. Die Fähigkeit zur Handlungsurheberschaft kann als Grundlage von Autonomie betrachtet werden, also dafür, dass handelnde Menschen ihre Handlungen nach Maximen ausrichten können, die sie sich selber setzen.

39) Die Umstände und Folgen von Handlungen können für deren moralische und rechtliche Bewertung von Bedeutung sein. So können aus Handlungen beispielsweise neben den beabsichtigten Folgen auch nicht beabsichtigte, aber den Handelnden erkennbare Folgen erwachsen. Dies ist relevant für das Konzept von Fahrlässigkeit, das im Kontext von KI eine große Rolle spielt. Und auch wenn primär einzelne Menschen handeln, schließt dies ein Konzept kollektiver Handlungen nicht aus, bei denen mehrere Personen von vornherein in einem Kontext der Koordination agieren.

40) Auch Technologie kann erheblichen Einfluss auf menschliches Handeln oder die menschliche Handlungserfahrung haben. Die zunehmende Durchdringung der menschlichen Lebenswelt mit informationstechnisch immer leistungsfähigeren Maschinen führt zu hybriden, soziotechnischen Konstellationen, in denen Menschen und Maschinen eng verwoben sind und auf komplexe Weise interagieren. Zudem können manche maschinellen Systeme menschliches Tun zum Teil so gut imitieren, dass die Simulation wie intentionales menschliches Handeln erscheint. Vor diesem Hintergrund ist es sinnvoll, an einem engen Handlungsbegriff, der an das zentrale Kriterium der Intentionalität gebunden ist, festzuhalten.

41) Das Intentionalitätskriterium ist zudem entscheidend für die Möglichkeit der Zuschreibung von Verantwortung im Kontext von Mensch-Maschine-Interaktionen

in zunehmend komplexer soziotechnischer Vernetzung. Verantwortung kann als Konzept einer fünffachen Relation gefasst werden: Wer (Verantwortungssubjekt) ist für was (Verantwortungsobjekt), gegenüber wem (Betroffenen), vor wem (Instanz) und unter welcher Norm verantwortlich?

42) In der Verantwortungsdiskussion zu wissenschaftlich-technischem Fortschritt ist zu bedenken, dass die Handlungsfolgen neuer Entwicklungen oft nur unter hohen und nicht eliminierbaren Wissensunsicherheiten abgeschätzt werden können. Verantwortungszuschreibung muss daher die Dimension des Handelns unter Unsicherheit berücksichtigen.

43) Moralische Verantwortung können nur natürliche Personen übernehmen, die über Handlungsfähigkeit verfügen, das heißt in der Lage sind, aktiv, zweckgerichtet und kontrolliert auf die Umwelt einzuwirken und dadurch Veränderungen zu verursachen. Träfe dies auch auf Maschinen zu, wären auch diese verantwortungsfähig. Dann müsste Maschinen der Personenstatus zugeschrieben werden, was jedoch weder aktuell noch angesichts der in absehbarer Zukunft erwartbaren qualitativen Entwicklungen maschineller Systeme angemessen wäre. Verantwortung kann daher nicht direkt von maschinellen Systemen übernommen werden, sondern nur von den Menschen, die in je unterschiedlichen Funktionen hinter diesen Systemen stehen, gegebenenfalls im Rahmen institutioneller Verantwortung.

44) Wer jeweils konkret wie viel Verantwortung trägt, ist häufig schwierig zu bestimmen. Die facettenreichen Verantwortungsgefüge zwischen Individuen, Organisationen und Staat werden noch komplexer, wenn die Wechselwirkungen zwischen diesen Beteiligten zumindest teilweise von algorithmischen Systemen gestützt oder vermittelt werden, welche mitunter kaum durchschaubar sind oder autonom zu agieren scheinen. Vor einem

solchen Hintergrund ist eine angemessene Gestaltung von Multiakteursverantwortung zentral.

45) Handlung, Vernunft und Verantwortung stehen im Zentrum humanistischer Philosophie. Menschen sind befähigt zur Handlungsurheberschaft und somit zur Autorschaft ihres Lebens. Sie sind frei und tragen daher Verantwortung für die Gestaltung ihres Handelns. Freiheit und Verantwortung sind zwei sich wechselseitig bedingende Aspekte menschlicher Autorschaft. Autorschaft ist wiederum an Vernunftfähigkeit gebunden.

46) Im Mittelpunkt dieser Trias aus Vernunft, Freiheit und Verantwortung steht das Phänomen der Affektion durch Gründe. Praktische Gründe sprechen für Handlungen, theoretische Gründe sprechen für Überzeugungen. In der Regel gibt es Gründe, das eine zu tun und das andere zu lassen, die gegeneinander abgewogen werden müssen. Der Konflikt von Gründen zwingt dann zur Abwägung und zur Systematisierung dieser Abwägung in Gestalt ethischer Theoriebildung.

47) Die menschliche Lebensform ist von reaktiven Einstellungen und moralischen Gefühlen geprägt, die von normativen Gründen begleitet sind. Freiheit kommt insofern ins Spiel, als wir diese zurückstellen, wenn wir erfahren, dass eine Person in ihrem Handeln nicht frei war. Diese Praxis der Zuschreibung von Freiheit und Verantwortung ist essenziell für die Grundlegung moralischer Beurteilung. Die Normen von Moral und Recht sind ohne die Annahme menschlicher Verantwortung und damit von Freiheits- und Vernunftfähigkeit unbegründet.

48) Eine Herausforderung dieser humanistischen Perspektive kommt aus den Neurowissenschaften. Empirische Studien, nach denen beispielsweise das motorische Zentrum des Gehirns schon mit der Vorbereitung einer Bewegung beginnt, bevor man sich bewusst für die Ausführung der Bewegung entschieden hat, werden mitunter als Beleg dafür interpretiert, dass es Freiheit und damit

menschliche Verantwortlichkeit nicht gebe. Tatsächlich lassen solche Befunde jedoch unterschiedliche Interpretationen zu und eignen sich nicht als Widerlegung menschlicher Freiheit und Verantwortung.

49) Eine zweite Kritik der humanistischen Anthropologie wird von der KI-Debatte inspiriert. Sie changiert zwischen einer Überwindung des Menschen in Gestalt des Transhumanismus, der mit neuen Mensch-Maschine-Symbiosen die Reichweite menschlichen Wirkens in neue Dimensionen heben möchte, und einem Maschinenparadigma, das den menschlichen Geist auf das Modell eines algorithmischen Systems reduziert. Gerade letzteres entfaltet besondere Relevanz im Kontext dieser Stellungnahme, da es großen Einfluss auf die Interpretation der Wechselwirkungen zwischen Mensch und Maschine und deren Rückwirkungen auf das menschliche Selbstverständnis hat.

50) In Maschinenparadigmen werden Menschen materialistisch als Maschinen oder Maschinen animistisch mit mentalen Zuständen ausgestattet und als menschengleich gedeutet. Die in KI-Diskursen teilweise verbreitete Tendenz, eine äußerliche Ununterscheidbarkeit von menschlicher und maschineller Performanz pauschal mit der Annahme von Intelligenz und Denkvermögen solcher Maschinen gleichzusetzen, ist das Ergebnis bestimmter theoretischer Vorannahmen insbesondere behavioristischer und funktionalistischer Art.

51) Der Behaviorismus versucht, menschliches Verhalten auf der Grundlage präzise beschreibbarer Reiz-Reaktion-Schemata zu erklären und die Psychologie damit in eine exakte Wissenschaft zu verwandeln; das Innenleben derart beschriebener Organismen wird dabei komplett ausgeblendet. Der Funktionalismus beruht auf der Annahme, dass mentale Zustände funktional vollständig erfasst werden können und die Frage nach der Seinsart mentaler Zustände zugunsten der genauen Beschreibung ihrer

Funktion aufgehoben werden kann und sollte. Durch die These der multiplen Realisierung, nach der bestimmte mentale Ereignisse, Eigenschaften oder Zustände durch ganz unterschiedliche physikalische Ereignisse, Eigenschaften oder Zustände realisiert werden können, scheint es zudem möglich, auch Computern mentale Zustände zuzuschreiben, obwohl sie keine biologischen Strukturen besitzen.

52) Kritik am Funktionalismus verweist auf phänomenales Bewusstsein, nach dem die mentalen Zustände eines Wesens entscheidend von Empfindungsqualitäten abhängen, die allein aufgrund äußeren Verhaltens nicht zugänglich sind. Dieses phänomenale Bewusstsein setzt dem Vermögen, die Qualität des Erlebens oder die mentalen Zustände anderer Lebewesen zu beurteilen, gewisse Grenzen und lässt die funktionalistisch inspirierte Mensch-Computer-Analogie als eine fragwürdige Reduktion erscheinen.

53) Ein weiteres Argument gegen den Funktionalismus stammt aus John Searles Gedankenexperiment zum „chinesischen Zimmer", in dem eine Person aus einer Kammer anhand einer genauen Gebrauchsanleitung chinesische Antworten auf Fragen herausreicht. Nicht diese Person beherrscht die chinesische Sprache und auch kein Übersetzungscomputer, sondern diejenigen, die die Gebrauchsanleitung bzw. den Algorithmus zur Beantwortung der Fragen verfasst haben.

54) In der Zurückweisung funktionalistischer Maschinenparadigmen wird die Bedeutung der gesamten Lebenserfahrung für die Vernunft deutlich. Menschliche Vernunft ist leibliche Vernunft. Der Leib ist Ausgangspunkt und Bestandteil jeder Wahrnehmung und Empfindung und als solcher Voraussetzung für menschliches In-der-Welt-Sein und die Herstellung von Beziehungen zu anderen. Kognitive Fähigkeiten sind in ihrem Entstehungs- und

Vollzugsprozess also an Sinnlichkeit und Leiblichkeit, Sozialität und Kulturalität gebunden.

55) Daraus ergeben sich auch Grenzen der Formalisierbarkeit und Simulierbarkeit menschlicher Vernunft. Die Aneignung menschlicher Erfahrung ist immer mit Deutungsprozessen verbunden und setzt immer ein Beteiligtsein, ein Engagement voraus. Auch hier spielt der Leib eine wichtige Rolle, denn er ermöglicht ein Handeln, das allein mittels bewusster Planung und Berechnung so nicht möglich wäre. Darin gründet eine Nichtsimulierbarkeit des Denkens, vor deren Hintergrund die Entwicklung von KI an Grenzen stößt.

56) Aus den bisherigen Überlegungen lässt sich zusammenfassen, dass menschliche Intelligenz unauflöslich mit den vielfältigen Dimensionen der menschlichen Lebenswelt verbunden ist. Sie operiert gründegeleitet und ist Ausdruck von akzeptierten Werten und Normen. Es ist fraglich, ob eine derart gründegeleitete, multidimensional bestimmte und soziokulturell eingebettete kohärente Praxis selbst für komplexe maschinelle Systeme jemals plausibel sein könnte.

Mensch-Technik-Relationen

57) Menschen entwickeln, gestalten und nutzen Technik als Mittel zum Zweck. Die mehr oder minder umfassende Delegation menschlicher Tätigkeiten an Maschinen – bis hin zur vollständigen Ersetzung menschlicher Handlungen durch maschinelle Vollzüge – wirkt allerdings häufig zurück auf menschliche Handlungsmöglichkeiten, Fertigkeiten, Autorschaft und Verantwortungsübernahme und kann diese jeweils erweitern oder vermindern. Die drei Begriffe des *Erweiterns*, *Verminderns* und *Ersetzens* dienen in dieser Stellungnahme als analytische Matrix.

58) Technikgestaltung wird im sozialen Konstruktivismus als Prozess beschrieben, der eher menschlich gesetzten, durch die jeweiligen gesellschaftlichen Prioritäten

geprägten Zwecken folgt. Im technologischen Determinismus wird eine insbesondere nach ökonomischen Verhältnissen bestimmende Eigendynamik als maßgeblich gesehen, der sich Mensch und Gesellschaft letztlich unterordnen und anpassen müssen. Tatsächlich spielen beide Ansätze zusammen und unterliegt die Mensch-Technik-Relation von Grund auf einem Ko-Konstruktionsverhältnis und kann als Ko-Evolution beschrieben werden. Soziale Kontexte und normative Kriterien auf der einen und Technologien auf der anderen Seite entwickeln sich weiter in gegenseitiger Wechselwirkung.

59) In ihrer Gesamtheit kann Technik dabei zu einer „zweiten Natur" werden, die Rand- und Erfolgsbedingungen für weiteres menschliches Leben setzt und auch die Weltsicht und das Problemlösen beeinflusst. Somit ist neue Technologie oft bereits das Ergebnis einer technologischen Art und Weise, wie Menschen die Welt sehen und sich zu ihr in Beziehung setzen. Die zunehmende Komplexität der Mensch-Technik- bzw. Mensch-Maschine-Relation verändert auch deren Wahrnehmung. In KI-gesteuerten Systemen scheinen die vormals klaren Unterscheidungen von Mensch und Technik weniger eindeutig zu werden. Auch in der Umgangssprache ist die Anthropomorphisierung digitaler Technik weit fortgeschritten, zum Beispiel in der Zuschreibung von Fähigkeiten wie Denken, Lernen, Entscheiden oder Zeigen von Emotion an KI und Roboter.

60) Subjekt-Objekt-Verhältnisse zwischen Mensch und Technik verändern sich ebenfalls. In vernetzten Systemen haben Menschen teils die Subjekt-, teils aber auch die Objektrolle inne. Wenn beispielsweise Entscheidungen über Menschen an Softwaresysteme delegiert werden, etwa hinsichtlich der Gewährung von Sozialleistungen, werden Menschen zu Objekten der „Entscheidungen" dieser Systeme, die hier auftreten, als ob sie Subjekte seien.

61) Verschiedene Ansätze versuchen, diese Entwicklungen in Konzepten zu mehrstufigen Mensch-Technik-Wechselwirkungen zu beschreiben.

62) Die Zuschreibung von Verantwortung bleibt in diesen Ansätzen jeweils beim Menschen. Moralisch problematische Resultate können dennoch durch KI-Systeme verursacht werden und sie haben Einfluss auf menschliches Handeln. Dieses ist also weder völlig autonom noch völlig sozial oder technisch determiniert, sondern in zunehmendem Maße soziotechnisch situiert.

63) KI zeigt in vielen Fällen eindeutig positive Folgen im Sinne der Erweiterung der Möglichkeiten menschlicher Autorschaft. Im Rahmen der Diffusion von Technik und Innovationen in die Gesellschaft, ihrer Nutzung und Veralltäglichung kommt es jedoch auch zu Verminderungen menschlicher Entfaltungsmöglichkeiten. Durch den Einsatz digitaler Technologien können Abhängigkeiten von diesen oder Anpassungsdruck entstehen – und andere, bis dahin etablierte Optionen verschlossen werden.

64) Solche Effekte können schleichend und teilweise unbewusst durch Verhaltensänderungen entstehen, ohne dass Intentionen von Akteuren dahinterstehen. Das Ersetzen als Endpunkt des Delegierens vormals menschlich ausgeübter Tätigkeiten an technische Systeme erfolgt jedoch intentional. Eine derartige Übertragung ist für sich genommen ein Ausdruck der Wahrnehmung menschlicher Autorschaft. Die zentrale ethische Frage ist, ob und wie diese Übertragung die Möglichkeiten *anderer* Menschen beeinflusst, vor allem jener, *über* die entschieden wird. Daraus ergibt sich ein Bedarf, die Übertragung menschlicher Tätigkeiten auf KI-Systeme auch gegenüber den davon Betroffenen transparent zu gestalten und bei der Beurteilung von KI zu berücksichtigen, *für wen* eine Anwendung jeweils Chancen oder Risiken und Erweiterungen oder Verminderungen der Autorschaft mit sich

bringt. Damit sind Aspekte sozialer Gerechtigkeit und Macht involviert.

65) Weiterhin sind psychologische Effekte im Zusammenhang mit KI-Systemen zu beachten, vor allem der Automation Bias. Menschen vertrauen algorithmisch erzeugten Ergebnissen und automatisierten Entscheidungsprozeduren häufig mehr als menschlichen Entscheidungen. Damit wird Verantwortung – zumindest unbewusst – an diese „Quasi-Akteure" delegiert. Selbst wenn ein KI-System normativ strikt auf die Rolle der Entscheidungsunterstützung begrenzt wird, kann Automation Bias dazu führen, dass ein KI-System allmählich in die Rolle des eigentlichen „Entscheiders" gerät und menschliche Autorschaft und Verantwortung ausgehöhlt werden.

TEIL II: AUSGEWÄHLTE ANWENDUNGEN UND SEKTORSPEZIFISCHE EMPFEHLUNGEN

Medizin

66) KI-gestützte digitale Produkte kommen zunehmend im Gesundheitssystem zum Einsatz. Die Betrachtung der mit ihnen verbundenen Chancen und Risiken bedarf einer wenigstens dreifachen Differenzierung. Erstens sind mehrere Akteursgruppen zu unterscheiden, die bezüglich eines KI-Einsatzes unterschiedliche Funktionen und Verantwortlichkeiten besitzen. Zweitens umfasst das Gesundheitswesen von der Forschung bis zur konkreten Patientenversorgung unterschiedliche Anwendungsbereiche für KI-Produkte. Drittens sind unterschiedliche Grade der Ersetzung menschlicher Handlungssegmente zu beobachten.

67) Bereits die Entwicklung geeigneter KI-Komponenten für die medizinische Praxis erfordert enge interdisziplinäre Zusammenarbeit verschiedener Sachverständiger und

stellt hohe Anforderungen an die Qualität der verwendeten Trainingsdaten, um vermeidbare Verzerrungen der Ergebnisse von vornherein zu minimieren. Systeme sind so zu konzipieren, dass sie Plausibilitätsprüfungen in der Nutzungsphase vorsehen, um den Gefahren eines Automation Bias zu entgehen. Mittels geeigneter Prüf-, Zertifizierungs- und Auditierungsmaßnahmen sollte gewährleistet werden, dass nur hinreichend geprüfte KI-Produkte zum Einsatz kommen, deren grundlegende Funktionsweise zumindest bei Systemen, die Entscheidungsvorschläge mit schwerwiegenden Konsequenzen für Betroffene unterbreiten, auch für diejenigen, die ein Produkt später verwenden, hinreichend erklär- sowie interpretierbar ist.

68) In der medizinischen Forschung kann der Einsatz von KI in mehrfacher Hinsicht vorteilhaft sein, sofern der Schutz der an den Studien teilnehmenden Personen und ihrer Daten gewährleistet ist. KI kann hier beispielsweise hilfreiche Vor- und Zuarbeiten bei Literaturrecherchen oder der Durchsuchung großer Datenbanken leisten, neue Korrelationen zwischen bestimmten Phänomenen entdecken und auf dieser Grundlage treffsichere Vorhersagen machen, etwa zur Ausbreitung eines Virus oder zur Struktur komplexer Moleküle.

69) In der medizinischen Versorgung werden KI-Instrumente zunehmend auch zur Diagnostik und Therapie eingesetzt, beispielsweise bei Brust- und Prostatakrebserkrankungen. Entscheidungsunterstützungssysteme modellieren und automatisieren hier Entscheidungsprozesse mittels Analyse verschiedener Parameter der Labordiagnostik, der Bildbearbeitung sowie der automatisierten Durchsicht von Patientenakten und wissenschaftlichen Datenbanken. Gerade Fortschritte in der KI-gestützten Bilderkennung eröffnen dabei neue Möglichkeiten einer frühzeitigen Detektion, Lokalisation und Charakterisierung pathologischer Veränderungen. In der Therapie

kommt KI beispielsweise in Operationsrobotern zum Einsatz.

70) Wenn ärztliche Tätigkeiten in derart engem bis mittleren Ausmaß an Technik delegiert werden, können beispielsweise Tumoren früher erkannt, Therapieoptionen erweitert und die Chancen auf eine erfolgreiche Therapie somit erhöht werden. Für ärztliches Personal eröffnet die Technik zudem die Chance, von monotonen Routinearbeiten entlastet zu werden und mehr Zeit für den Austausch mit Patientinnen und Patienten zu gewinnen. Diesen Chancen stehen aber auch Risiken gegenüber, beispielweise wenn Fachkräfte durch die fortschreitende Delegation bestimmter Aufgaben an technische Systeme eigene Kompetenzen verlieren oder Sorgfaltspflichten im Umgang mit KI-gestützter Technik aufgrund eines Automation Bias vernachlässigen.

71) Um die Chancen des KI-Einsatzes in klinischen Situationen zu realisieren und Risiken zu minimieren, sind mehrere Ebenen zu berücksichtigen. So bedarf es unter anderem einer flächendeckenden und möglichst einheitlichen technischen Ausrüstung, Personalschulung und kontinuierlichen Qualitätssicherung ebenso wie Strategien, die gewährleisten, dass auch in KI-gestützten Verfahren Befunde auf Plausibilität geprüft werden, die persönliche Lebenssituation von Erkrankten umfassend berücksichtigt und vertrauensvoll kommuniziert wird. Auch der bei den meisten medizinischen KI-Anwendungen große Datenbedarf bringt Herausforderungen mit sich, sowohl hinsichtlich des Schutzes der Privatsphäre Betroffener als auch mit Blick auf eine teils sehr restriktive individuelle Auslegungspraxis geltender Datenschutzbestimmungen, die der Realisierung von Potenzialen des KI-Einsatzes in der klinischen Praxis im Wege stehen kann.

72) Einer der wenigen medizinischen Handlungsbereiche, in denen KI-basierte Systeme zum Teil bereits ärztliches bzw. anderes Gesundheitspersonal – jedenfalls de facto

– weitgehend oder vollständig ersetzen, ist die Psychotherapie. Hier sind seit einigen Jahren Instrumente in Entwicklung und Nutzung, meist in Form von bildschirmbasierten Apps, die auf algorithmischer Basis eine Art von Therapie anbieten und vielfach frei erhältlich sind. Solche Apps können einerseits angesichts ihrer Niedrigschwelligkeit und ständigen Verfügbarkeit Menschen in Erstkontakt mit therapeutischen Angeboten bringen, die sonst zu spät oder gar nicht eine Therapie erhalten. Andererseits gibt es Bedenken etwa hinsichtlich mangelnder Qualitätskontrollen, dem Schutz der Privatsphäre oder wenn Menschen eine Art emotionale Beziehung zur therapeutischen App aufbauen. Kontrovers diskutiert wird auch, ob die zunehmende Nutzung solcher Apps weiterem Abbau von therapeutischem Fachpersonal Vorschub leistet.

73) Auf Grundlage dieser Überlegungen formuliert der Deutsche Ethikrat neun Empfehlungen für den Einsatz von KI im Gesundheitssektor:

- *Empfehlung Medizin 1:* Bei der Entwicklung, Erprobung und Zertifizierung medizinischer KI-Produkte bedarf es einer engen Zusammenarbeit mit den relevanten Zulassungsbehörden sowie insbesondere mit den jeweils zuständigen medizinischen Fachgesellschaften, um Schwachstellen der Produkte frühzeitig zu entdecken und hohe Qualitätsstandards zu etablieren.
- *Empfehlung Medizin 2:* Bei der Auswahl der Trainings-, Validierungs- und Testdatensätze sollte über bestehende Rechtsvorgaben hinaus mit einem entsprechenden Monitoring sowie präzise und zugleich sinnvoll umsetzbaren Dokumentationspflichten sichergestellt werden, dass die für die betreffenden Patientengruppen relevanten Faktoren (z. B. Alter, Geschlecht, ethnische Einflussfaktoren,

Vorerkrankungen und Komorbiditäten) hinreichend berücksichtigt werden.

- *Empfehlung Medizin 3:* Bei der Gestaltung des Designs von KI-Produkten zur Entscheidungsunterstützung ist sicherzustellen, dass die Ergebnisdarstellung in einer Form geschieht, die Gefahren etwa von Automatismen (Automation Bias) transparent macht, ihnen entgegenwirkt und die die Notwendigkeit einer reflexiven Plausibilitätsprüfung der jeweils vom KI-System vorgeschlagenen Handlungsweise unterstreicht.
- *Empfehlung Medizin 4:* Bei der Sammlung, Verarbeitung und Weitergabe von gesundheitsbezogenen Daten sind generell strenge Anforderungen und hohe Standards in Bezug auf Aufklärung, Datenschutz und Schutz der Privatheit zu beachten. In diesem Zusammenhang verweist der Deutsche Ethikrat auf seine 2017 im Kontext von Big Data und Gesundheit formulierten Empfehlungen, die sich am Konzept der Datensouveränität orientieren, das für den Bereich von KI-Anwendungen im Gesundheitsbereich gleichermaßen Gültigkeit entfaltet.
- *Empfehlung Medizin 5:* Bei durch empirische Studien sorgfältig belegter Überlegenheit von KI-Anwendungen gegenüber herkömmlichen Behandlungsmethoden ist sicherzustellen, dass diese allen einschlägigen Patientengruppen zur Verfügung stehen.
- *Empfehlung Medizin 6:* Für erwiesen überlegene KI-Anwendungen sollte eine rasche Integration in die klinische Ausbildung des ärztlichen Fachpersonals erfolgen, um eine breitere Nutzung vorzubereiten und verantwortlich so gestalten zu können, dass möglichst alle Patientinnen und Patienten davon profitieren und bestehende Zugangsbarrieren zu den neuen Behandlungsformen abgebaut werden. Dazu ist die Entwicklung einschlägiger Curricula/Module in Aus-, Fort- und Weiterbildung notwendig. Auch

die anderen Gesundheitsberufe sollten entsprechende Elemente in die Ausbildung aufnehmen, um die Anwendungskompetenz bei KI-Anwendungen im Gesundheitsbereich zu stärken.

» *Empfehlung Medizin 7:* Bei routinemäßiger Anwendung von KI-Komponenten sollte nicht nur gewährleistet werden, dass bei denjenigen, die sie klinisch nutzen, eine hohe methodische Expertise zur Einordnung der Ergebnisse vorhanden ist, sondern auch strenge Sorgfaltspflichten bei der Datenerhebung und -weitergabe sowie bei der Plausibilitätsprüfung der maschinell gegebenen Handlungsempfehlungen eingehalten werden. Besondere Aufmerksamkeit erfordert die Gefahr eines Verlustes von theoretischem wie haptisch-praktischem Erfahrungswissen und entsprechenden Fähigkeiten (Deskilling); dieser Gefahr sollte mit geeigneten, spezifischen Fortbildungsmaßnahmen entgegengewirkt werden.

» *Empfehlung Medizin 8:* Bei fortschreitender Ersetzung ärztlicher, therapeutischer und pflegerischer Handlungssegmente durch KI-Komponenten ist nicht nur sicherzustellen, dass Patientinnen und Patienten über alle entscheidungsrelevanten Umstände ihrer Behandlung vorab informiert werden. Darüber hinaus sollten auch gezielte kommunikative Maßnahmen ergriffen werden, um dem drohenden Gefühl einer zunehmenden Verobjektivierung aktiv entgegenzuwirken und das Vertrauensverhältnis zwischen den beteiligten Personen zu schützen. Je höher der Grad der technischen Substitution menschlicher Handlungen durch KI-Komponenten ist, desto stärker wächst der Aufklärungs- und Begleitungsbedarf der Patientinnen und Patienten. Die verstärkte Nutzung von KI-Komponenten in der Versorgung darf nicht zu einer weiteren Abwertung der sprechenden Medizin oder einem Abbau von Personal führen.

» *Empfehlung Medizin 9:* Eine vollständige Ersetzung der ärztlichen Fachkraft durch ein KI-System gefährdet das Patientenwohl und ist auch nicht dadurch zu rechtfertigen, dass schon heute in bestimmten Versorgungsbereichen ein akuter Personalmangel besteht. Gerade in komplexen Behandlungssituationen bedarf es eines personalen Gegenübers, das durch technische Komponenten zwar immer stärker unterstützt werden kann, dadurch selbst als Verantwortungsträger für die Planung, Durchführung und Überwachung des Behandlungsprozesses aber nicht überflüssig wird.

Bildung

74) Auch in der schulischen Bildung kommen zunehmend digitale Technologien und algorithmische Systeme zum Einsatz. Dies kann sowohl zur Standardisierung von Lernprozessen führen als auch mehr Personalisierung ermöglichen. Die Einsatzmöglichkeiten reichen von sehr eng umrissenen punktuellen Angeboten bis hin zu Szenarien, in denen KI-gestützte Lehr- und Lernsysteme zeitweise oder gänzlich eine Lehrkraft ersetzen können.

75) Das hier zugrunde gelegte Verständnis von Bildung orientiert sich an der Fähigkeit des Menschen zu freiem und vernünftigem Handeln, das nicht auf behavioristische oder funktionalistische Modelle zu reduzieren ist. Bildung erfordert den Erwerb von Orientierungswissen als Bedingung von reflexiver Urteilskraft und Entscheidungsstärke. Dieser Prozess umfasst auch kulturelles Lernen sowie emotionale und motivationale Aspekte. Das Lehr- und Lerngeschehen ist als dynamische Interaktion mit anderen Personen zu begreifen. Der Einsatz von KI-gestützten Instrumenten in der Schule ist daraufhin zu überprüfen, ob er einem solchen Verständnis des Menschen als einer zur Selbstbestimmung und

Verantwortung fähigen Person entspricht und solche Prozesse fördert oder ob er diesen entgegensteht.

76) Ausgangspunkt der meisten KI-Anwendungen in der Bildung ist die Sammlung und Auswertung vieler Daten der Lernenden und mitunter auch der Lehrkräfte. Hier stellen sich Fragen nach dem sinnvollen Grad und Ausmaß der Datenerhebung sowie deren wünschenswerten Verwendungsweisen. Es geht darum, Lernende in ihrem individuellen Lernprozess durch Datennutzung bestmöglich zu unterstützen und gleichzeitig zu verhindern, dass diese Daten zur Überwachung oder Stigmatisierung von einzelnen Lernenden missbraucht werden können.

77) Auf Grundlage der erhobenen Daten können individualisierte Rückmeldungen über Lehr- und Lernprozesse sowie entsprechende Reaktionen oder Empfehlungen des Softwaresystems erfolgen. Durch Auswertung von zum Beispiel Lerngeschwindigkeit, typischen Fehlern, Stärken und Schwächen kann die Software das Lernprofil der Lernenden erkennen und die Lerninhalte entsprechend anpassen. Subjektive Eindrücke der Lehrkräfte können dadurch datenbasiert untermauert, aber auch korrigiert werden.

78) Auch in der Schule kann es durch KI zu engen, mittleren und weiten Ersetzungen bestimmter Handlungssegmente und Interaktionen kommen. Eine enge Ersetzung liegt etwa vor, wenn ein Softwaresystem für einen genau bestimmten Lernabschnitt eingesetzt wird. Aufwändigere und datenintensivere intelligente Tutorsysteme können auch komplexere Lerninhalte in unterschiedlichen Fächern im Zusammenwirken mit Lernenden vermitteln und so breitere Teilaspekte des Unterrichtsgeschehens ersetzen oder im Einzelfall die Funktion einer Lehrkraft vollständig übernehmen.

79) Darüber hinaus gibt es mittlerweile auch Bestrebungen, KI zur Analyse des Verhaltens im Klassenraum einzusetzen (Classroom Analytics), um die Dynamik ganzer

Lerngruppen umfassend zu dokumentieren und auszuwerten. Solche Ansätze sind aufgrund der für sie notwendigen Erfassung vielfältiger Daten unter anderem über das Verhalten von Schülerinnen und Schülern sowie Lehrkräften umstritten. Chancen auf verbesserte Pädagogik und Didaktik stehen potenziell negative Auswirkungen umfangreicher Datensammlungen auf die Privatsphäre und Autonomie aller Beteiligten gegenüber.

80) Ein besonders kontrovers diskutierter Aspekt von Classroom Analytics betrifft die mögliche Erfassung von Aufmerksamkeit (Attention Monitoring) oder emotionaler Verfasstheit (Affect Recognition) der im Unterricht interagierenden Personen, insbesondere basierend auf der Analyse von Audio- oder Videodaten aus Klassenräumen. Auch wenn dies durchaus mit dem Ziel einer Verbesserung von Lernergebnissen verbunden sein kann, wird bezweifelt, dass Aufmerksamkeit und Emotionen jedenfalls mit aktueller Technik hinreichend genau, zuverlässig und ohne systematische Verzerrung gemessen werden können. Außerdem werden die vorstehend genannten Risiken der notwendigen Datenerfassung hier als besonders gravierend eingeschätzt.

81) Zusammenfassend lassen sich aufseiten der Chancen von KI in der Schule personalisiertes Lernen und Entlastung von Lehrkräften anführen, ebenso eine potenziell objektivere und fairere Bewertung von Lernergebnissen sowie mitunter verbesserte Zugangschancen und Möglichkeiten zur Inklusion von Lernenden mit besonderen Bedürfnissen. Zu den Risiken gehören neben den bereits erwähnten Bedenken hinsichtlich Verzerrungen und Beeinträchtigungen der Privatsphäre und der Autonomie auch Gefahren der Isolation und Vereinsamung von Lernenden sowie möglicherweise qualitative Veränderungen des Lernverhaltens. So könnten sich etwa grundsätzliche Auswirkungen auf die Motivation und Fähigkeit

von Schülerinnen und Schülern zur Lösung komplexerer Aufgaben ergeben.

82) KI-gestützte Lehr- und Lernsysteme können den jeweiligen Lernprozess unterstützen, ersetzen aber nicht die personale Vermittlung und die personalen Aspekte von Bildung. Die Relevanz der Schule als Sozialraum der Interaktion zwischen Menschen ist dabei nicht zu unterschätzen. Da Bildung nicht nur in optimierbarer und berechenbarer Anhäufung von Wissen, sondern vor allem in einem konstruktiven und verantwortlichen Umgang mit erlerntem Wissen besteht, ist bei der Delegation von Elementen des Lehr- und Lerngeschehens an Maschinen besonders darauf zu achten, dass Lernprozesse, die zentral für die Persönlichkeitsbildung des Menschen sind, dadurch nicht vermindert werden.

83) Vor dem Hintergrund dieser Überlegungen legt der Deutsche Ethikrat elf Empfehlungen für den Einsatz von KI in der schulischen Bildung vor:

- *Empfehlung Bildung 1:* Digitalisierung ist kein Selbstzweck. Der Einsatz sollte nicht von technologischen Visionen, sondern von grundlegenden Vorstellungen von Bildung, die auch die Bildung der Persönlichkeit umfassen, geleitet sein. Die vorgestellten Tools sollten deshalb im Bildungsprozess kontrolliert und als ein Element innerhalb der Beziehung zwischen Lehrenden und Lernenden eingesetzt werden.
- *Empfehlung Bildung 2:* Für jedes Einsatzgebiet gilt es, eine angemessene Abwägung von Chancen und Risiken vorzunehmen. Insbesondere sollten Autonomie und Privatheit von Lehrenden und Lernenden hohen Schutz erfahren. Besondere Chancen ergeben sich im Bereich der Inklusion und Teilhabe, wo das Potenzial dieser Systeme genutzt werden sollte, um etwa sprachliche oder räumliche Barrieren abzubauen.

- *Empfehlung Bildung 3:* Tools, die einzelne Elemente des Lehr- und Lernprozesses ersetzen bzw. ergänzen (enge Ersetzung) und nachweislich Fähigkeiten, Kompetenzen oder soziale Interaktion der Personen, die sie nutzen, erweitern, wie etwa einige intelligente Tutorsysteme oder Telepräsenzroboter für externe Lehrbeteiligung, sind prinzipiell weniger problematisch als solche, die umfassendere bzw. weitere Teile des Bildungsprozesses ersetzen. Je höher der Ersetzungsgrad, desto strenger müssen Einsatzbereiche, Umgebungsfaktoren und Nutzenpotenziale evaluiert werden.
- *Empfehlung Bildung 4:* Es gilt, standardisierte Zertifizierungssysteme zu entwickeln, die anhand transparenter Kriterien des Gelingens von Lernprozessen im genannten umfassenden Sinne Schulämter, Schulen und Lehrkräfte dabei unterstützen können, sich für oder gegen die Nutzung eines Produkts zu entscheiden. Hier kann sich auch der Empfehlung zur dauerhaften Einrichtung länderübergreifender Zentren für digitale Bildung, wie es im Gutachten „Digitalisierung im Bildungssystem: Handlungsempfehlungen von der Kita bis zur Hochschule" der Ständigen Wissenschaftlichen Kommission der Kultusministerkonferenz angesprochen wurde, angeschlossen werden.
- *Empfehlung Bildung 5:* Bei der Entwicklung, Erprobung und Zertifizierung entsprechender KI-Produkte bedarf es einer engen Zusammenarbeit mit den relevanten Behörden, mit den jeweils zuständigen pädagogischen Fachgesellschaften sowie der Partizipation von Beteiligten, um Schwachstellen der Produkte frühzeitig zu entdecken und hohe Qualitätsstandards zu etablieren. Bekannte Herausforderungen KI-getriebener Technologien wie beispielsweise Verzerrungen bzw. Bias oder Anthropomorphisierungstendenzen

sollten bei der Entwicklung und Standardisierung berücksichtigt werden.

- *Empfehlung Bildung 6:* Um den verantwortlichen Einsatz von KI-Technologien im Bildungsprozess zu gewährleisten, muss die Nutzungskompetenz insbesondere der Lehrkräfte erhöht werden; es bedarf der Entwicklung und Etablierung entsprechender Module und Curricula in der Aus-, Fort- und Weiterbildung. Insbesondere die Gefahren eines verengten pädagogischen Ansatzes und eines Deskillings in der Lehre sollten dabei aktiv in den Blick genommen werden. Ebenso sollte die digitale Nutzungskompetenz von Lernenden sowie Eltern gestärkt und um KI-Aspekte erweitert werden.
- *Empfehlung Bildung 7:* Im Sinne der Beteiligungsgerechtigkeit sollten KI-basierte Tools Lernenden grundsätzlich auch für das Eigenstudium zur Verfügung stehen.
- *Empfehlung Bildung 8:* Die Einführung von KI-Tools im Bildungsbereich erfordert ferner den Ausbau verschiedener flankierender Forschungsbereiche. Sowohl theoretische Fundierung als auch empirische Evidenz zu Effekten, etwa auf die Kompetenzentwicklung (z. B. Problemlösen), oder zur Beeinflussung der Persönlichkeitsentwicklung bei Kindern und Heranwachsenden muss weiter ausgebaut werden. Dabei sollte nicht nur stärker in Forschung und entsprechende Produktentwicklung investiert, sondern vor allen Dingen auch die praktische Erprobung und Evaluation im schulischen Alltag verstärkt werden.
- *Empfehlung Bildung 9:* Des Weiteren stellt sich hier die Problematik der Datensouveränität. Zum einen sind bei der Sammlung, Verarbeitung und Weitergabe von bildungsbezogenen Daten strenge Anforderungen an den Schutz der Privatsphäre zu beachten. Zum anderen sollte die gemeinwohlorientierte,

verantwortliche Sammlung und Nutzung von großen Daten, etwa in der prognostischen lehrunterstützenden Anwendung, ermöglicht werden.

» *Empfehlung Bildung 10:* Eine vollständige Ersetzung von Lehrkräften läuft dem hier skizzierten Verständnis von Bildung zuwider und ist auch nicht dadurch zu rechtfertigen, dass schon heute in bestimmten Bereichen ein akuter Personalmangel und eine schlechte (Aus-)Bildungssituation herrschen. In der komplexen Situation der schulischen Bildung bedarf es eines personalen Gegenübers, das mithilfe technischer Komponenten zwar immer stärker unterstützt werden kann, dadurch selbst als Verantwortungsträger für die pädagogische Begleitung und Evaluation des Bildungsprozesses aber nicht überflüssig wird.

» *Empfehlung Bildung 11*: In Anbetracht der erkenntnistheoretischen und ethischen Herausforderungen und unter Abwägung potenzieller Nutzen und Schäden stehen die Mitglieder des Deutschen Ethikrates dem Einsatz von Audio- und Videomonitoring im Klassenzimmer insgesamt skeptisch gegenüber. Insbesondere erscheint die Analyse von Aufmerksamkeit und Emotionen per Audio- und Videoüberwachung des Klassenraums mittels aktuell verfügbarer Technologien nicht vertretbar. Ein Teil des Ethikrates schließt den Einsatz von Technologien zur Aufmerksamkeits- und Affekterkennung zukünftig jedoch nicht vollständig aus, sofern sichergestellt ist, dass die erfassten Daten eine wissenschaftlich nachweisbare Verbesserung des Lernprozesses bieten und das hierfür notwendige Monitoring von Lernenden und Lehrkräften keine inakzeptablen Auswirkungen auf deren Privatsphäre und Autonomie hat. Ein anderer Teil des Ethikrates hingegen hält die Auswirkungen auf Privatsphäre, Autonomie und Gerechtigkeit hingegen generell für nicht akzeptabel und befürwortet

daher ein Verbot von Technologien zu Aufmerksamkeitsmonitoring und Affekterkennung in Schulen.

Öffentliche Kommunikation und Meinungsbildung

84) Durch die digitale Transformation verändern sich auch politisch relevante Kommunikationsprozesse. Die rasante Verbreitung digitaler Plattformen und sozialer Medien mit ihren algorithmisch vermittelten Informations- und Kommunikationsangeboten wirkt sich nicht nur auf einzelne gesellschaftliche Sphären aus, sondern potenziell auch auf große Teile der öffentlichen Kommunikation und Meinungsbildung – mit Konsequenzen für das demokratische Legitimationsgefüge.

85) Viele Plattformen bieten inzwischen sich ähnelnde Möglichkeiten an, multimediale Inhalte zu erstellen und zu verbreiten, auf die Inhalte anderer zu reagieren, sich mit anderen Personen auszutauschen und die Plattform nach Inhalten zu durchsuchen oder diese zu abonnieren. Auch Optionen, eigene Inhalte gezielt zu bewerben und Produkte und Dienstleistungen direkt anzubieten oder zu kaufen, sind vielfach vorhanden. Fast alle weiter verbreiteten Plattformen und Dienste werden von privaten Unternehmen aus den USA oder China betrieben und die größten sozialen Netzwerke gehören nur wenigen Firmen. Aufgrund dieser Marktmacht sowie der Vielseitigkeit und Integration der Dienste funktionieren viele Angebote inzwischen als reichhaltige soziotechnische Infrastrukturen, in denen sich ein Großteil des Online-Nutzungsverhaltens nach den Vorgaben weniger Konzerne abspielt.

86) Mit der Fülle der Informationen und Interaktionsmöglichkeiten in sozialen Medien gehen technische Herausforderungen und ökonomische Potenziale einher, die gemeinsam zur Gestaltung aktueller Funktionsweisen und Geschäftsmodelle beigetragen haben. Die Fülle der Inhalte stellt Plattformen wie auch Kundschaft vor das

Problem der Informationsauswahl. Diese wird aktuell überwiegend an Algorithmen delegiert, die dafür sorgen, dass jeder Person beim Besuch einer Plattform auf sie persönlich zugeschnittene Inhalte in einer bestimmten Reihenfolge angezeigt werden.

87) Die Kriterien, nach denen solche Algorithmen ihre Auswahl treffen, sind eng mit ökonomischen Faktoren verknüpft. Die meisten Plattformen und Dienste folgen einem werbebasierten Geschäftsmodell, das am besten funktioniert, wenn die Interessen der einzelnen Nutzerinnen und Nutzer erstens möglichst präzise bekannt sind und Menschen zweitens möglichst viel Zeit auf der Plattform verbringen, während derer ihnen auf persönliche Interessen zugeschnittene Werbung präsentiert wird. Daher lohnt es sich für Plattformen, so viele Datenspuren wie möglich über den persönlichen Hintergrund, die Interessen, das Nutzungsverhalten und das soziale Netzwerk der Personen, die sie nutzen, zu sammeln und für die Auswahl personalisierter Inhalte zu verwenden (Profiling).

88) Eine algorithmisch gesteuerte personalisierte Informationsauswahl, in der ökonomische und aufmerksamkeitsbasierte Faktoren derart eng verbunden sind und die sich anhand des Nutzungsverhaltens ständig weiterentwickelt, führt dazu, dass Inhalte, die besonders sensationell erscheinen oder intensive emotionale Reaktionen auslösen, sich überproportional schnell und weit verbreiten. Dies begünstigt unter anderem Falschnachrichten und Inhalte wie Hassrede, Beleidigungen und Volksverhetzungen.

89) In Reaktion auf die Herausforderung, wie mit solchen potenziell problematischen und gleichzeitig verbreitungsstarken Inhalten umzugehen ist, bemühen sich Plattformen darum, ihre Inhalte nach verschiedenen Kriterien zu moderieren (Content-Moderation). Hierbei kommen sowohl Menschen als auch algorithmische

Systeme zum Einsatz. Grundlage für die Moderation sind rechtliche Vorgaben sowie plattformeigene Kommunikationsregeln, auf deren Grundlage auch rechtlich zulässige Inhalte gelöscht, gesperrt oder in ihrer Reichweite eingeschränkt werden können.

90) Menschliche Moderation erfolgt typischerweise durch Personen, die bei Drittanbietern angestellt sind, mit denen eine Plattform vertraglich zusammenarbeitet. Diese Personen werden bei oftmals prekären Arbeitsbedingungen mit häufig extrem belastendem Material wie Tötungen, Kindesmissbrauch, Tierquälerei und Suizid konfrontiert. Zudem müssen sie innerhalb weniger Sekunden sprachlich und kulturell komplexe Nuancen berücksichtigen, von denen die Zulässigkeit eines Beitrags entscheidend abhängen kann.

91) Algorithmen können im Gegensatz dazu anstößige Inhalte herausfiltern, ohne dass diese durch Menschen angesehen werden müssen, und darüber hinaus mit der unübersichtlichen Menge an Daten und Inhalten im Netz besser umgehen. Allerdings sind automatisierte Methoden jedenfalls bislang häufig unzureichend, um den kulturellen und sozialen Zusammenhang einer Äußerung einzubeziehen und diese damit adäquat zu beurteilen. Aufgrund der aktuellen rechtlichen Anreizstruktur besteht die Gefahr, dass systematisch auch Inhalte gelöscht oder unzugänglich gemacht werden, die nicht gegen Regeln verstoßen (Overblocking).

92) Durch die beschriebenen Funktionsweisen von Plattformen und die sich dabei entfaltenden soziotechnischen Verquickungen können menschliche Handlungsfähigkeiten in unterschiedlicher Weise erweitert oder vermindert werden. Die Delegation von Kuratierungs- und Moderationsprozessen an Algorithmen ist mit Komfort- und Effizienzgewinnen verbunden und kann eine Erweiterung von Handlungsmöglichkeiten bedeuten, wenn beispielsweise Informationen und persönliche Ziele

besser oder schneller erreicht werden können oder wenn aufgrund der effektiven Delegation der Inhaltsauswahl an Algorithmen Entlastungseffekte auftreten, die Freiräume für andere Aktivitäten schaffen.

93) Eine Verminderung menschlicher Handlungsspielräume und persönlicher Freiheit kann sich ergeben, wo es Menschen schwerfällt, sich dem Sog von Plattformangeboten zu entziehen und ihre Nutzung dieser Angebote auf ein für sie gesundes Maß zu beschränken. Zudem kann eine algorithmische Kuratierung Autorschaft vermindern, wenn eine rationale Auseinandersetzung mit Alternativen durch die algorithmische Vorwegnahme bestimmter Relevanzentscheidungen nur noch eingeschränkt stattfinden kann.

94) Neben diesen allgemeinen Auswirkungen der Funktionsweisen von Plattformen und sozialen Medien verändern sich durch sie auch die Informations- und die Diskursqualität, welche wichtige Grundlagen der öffentlichen Meinungsbildung sind – mit potenziell weitreichenden Konsequenzen für Prozesse der politischen Willensbildung. Wie weit verbreitet und wirkmächtig die nachfolgend genannten Effekte sind, lässt sich aktuell zwar noch nicht abschließend beurteilen, auch weil die Datenlage mitunter unklar oder widersprüchlich ist. Ein genauerer Blick auf die postulierten Mechanismen lohnt jedoch schon deswegen, weil die von ihnen berührten Prozesse grundlegend für unsere Demokratie sind.

95) Mit Blick auf die Informationsqualität ist zunächst auf die positive Erweiterung vieler Informationsmöglichkeiten zu verweisen. Demgegenüber wird vielfach die Sorge geäußert, dass die derzeit gängigen Praktiken algorithmischer Kuratierung auch negative Auswirkungen haben, indem sie die Verbreitung von Falschnachrichten und Verschwörungstheorien fördern, zur Entstehung von Filterblasen und Echokammern beitragen und Inhalte

priorisieren, die negative emotionale und moralische Reaktionen und Interaktionen provozieren.

96) Trotz der genannten Unsicherheiten über das Ausmaß der beschriebenen Effekte erscheint es plausibel, dass Falschnachrichten, Filterblasen und Echokammern sowie eine emotional-moralische Zuspitzung vieler Inhalte negative Auswirkung auf die Informationsqualität entfalten können. Die Freiheit, qualitativ hochwertige Informationen zu finden, wird unter diesen Umständen durch die Wirkmacht der zum Einsatz kommenden Algorithmen praktisch vermindert.

97) Änderungen in der Qualität, Darbietung und Verbreitung algorithmisch vermittelter Informationen betreffen auch die Diskursqualität in ethisch wie politisch relevanter Hinsicht. Auch hier sind zunächst wieder positive Entwicklungen und Potenziale zu benennen, die sich insbesondere aus den auf Plattformen und in sozialen Medien wesentlich erhöhten Möglichkeiten zu Teilhabe und direkter Vernetzung ergeben. Gegenüber den genannten Chancen werden jedoch auch mit Blick auf die Diskursqualität negative Entwicklungen diskutiert. Dabei geht es vor allem um drei Themen: politische Polarisierung öffentlicher Diskurse, politische Werbung und Manipulation sowie das Spannungsfeld von Diskursverrohungen und überbordenden Eingriffen in die Äußerungs- und Meinungsfreiheit.

98) Es gibt viele Hinweise, dass die beschriebene Verbreitungsfähigkeit emotional und moralisch aufgeladener Inhalte zu Tonfallverschiebungen geführt hat, auch und insbesondere auf Kanälen, die aktiv zur Gestaltung des politischen Diskurses beitragen. Nachdem beispielsweise geänderte Auswahlkriterien auf Facebook dazu führten, dass sich künftig vor allem solche Inhalte erfolgreich verbreiteten, auf die Menschen besonders aufgebracht reagieren, verschärften viele politische

Kommunikationsteams den Tonfall ihrer Beiträge, um diesen Kriterien gerecht zu werden.

99) Auf Plattformen ergibt sich zudem viel Potenzial für besonders wirkmächtige Kommunikationskampagnen, die eingebettet in den digitalen Alltag ablaufen, ohne dass Nutzerinnen und Nutzer sich dessen gewahr werden. Die reichhaltigen datenbasierten Profile, die sich aus dem Nutzungsverhalten auf Plattformen erstellen lassen, können auch genutzt werden, um zielgenaue politische Werbung zu schalten (Targeted Advertisement) oder um Menschen strategisch zu desinformieren oder von der Wahl abzuhalten. Wie erfolgreich solche auch als Microtargeting bezeichneten Ansätze sind, ist zwar noch nicht hinreichend erforscht, doch allein das Wissen darum, dass versucht wird, auf Grundlage sehr persönlicher psychologischer Merkmale politische Präferenzen zu manipulieren, kann plausibel negative Effekte auf den politischen Diskurs und das Vertrauen in politische Meinungsbildungsprozesse entfalten.

100) Vertrauensschädigend kann sich weiterhin der Umstand auswirken, dass zur strategischen Beeinflussung des öffentlichen politischen Diskurses auch vielfach unechte Profile (Fake-Accounts) eingesetzt werden, die teilweise automatisiert betriebenen werden (Bots). Kommunikationskampagnen, die solche unechten Profile nutzen, können damit Botschaften effektiv verstärken, ihnen somit mehr Überzeugungskraft verleihen und Diskurse mitunter problematisch verzerren.

101) Der bereits beschriebene Trend zur Verschärfung von Tonlagen auf Plattformen und in sozialen Medien geht mit der Sorge einher, dass eine Zunahme stark negativ und aggressiv geprägter Kommunikationsstile bis hin zu Hassrede, Drohungen und Gewaltaufforderungen zu einer Verrohung des politischen Diskurses beitragen kann. Selbst wenn online verbreitete Hetze nicht in Handlungen in der realen Welt umschlägt, kann es auch hier zu

Chilling-Effekten kommen. Wo nämlich solche Äußerungen so viel Unbehagen und Angst schüren, dass dies Personen davon abhält, sich am öffentlichen Diskurs zu beteiligen, wirkt dies auf die Freiheit und Handlungsmöglichkeiten Betroffener in der Onlinekommunikation vermindernd.

102) Andererseits werden auch Bemühungen, potenziell problematische Inhalte mit Moderationsmaßnahmen einzudämmen, teils kritisch beurteilt, denn solche Reaktionen werfen ihrerseits demokratietheoretische Fragen auf. Übermäßige Löschungen und Sperrungen können einen Eingriff in die Meinungs- und Pressefreiheit darstellen und selbst zu Chilling-Effekten beitragen, nämlich dann, wenn Menschen bestimmte Inhalte gar nicht erst veröffentlichen, weil sie befürchten, dass diese Inhalte gleich wieder gelöscht oder gar ihre Accounts (zeitweise) gesperrt werden könnten.

103) In der Zusammenschau können die hier aufgezeigten Phänomene und Entwicklungen, die sich in den soziotechnischen Infrastrukturen digitaler Netzwerke vollziehen, erhebliche Auswirkungen auf Prozesse der öffentlichen Kommunikation sowie der politischen Meinungs- und Willensbildung entfalten, auch und vielleicht insbesondere in demokratischen Gesellschaften. Vor diesem Hintergrund legt der Deutsche Ethikrat zu diesem Anwendungsfeld von KI zehn Empfehlungen vor:

» *Empfehlung Kommunikation 1:* Regulierung sozialer Medien: Es bedarf klarer rechtlicher Vorgaben, in welcher Form und in welchem Ausmaß soziale Medien und Plattformen über ihre Funktions- und Vorgehensweisen zur Kuratierung und Moderation von Inhalten informieren müssen und wie dies auf der Grundlage institutioneller Regelungen umgesetzt wird. Dies muss durch externe Kontrollen überprüfbar sein; rein freiwillige Ansätze privater Handelnder,

insbesondere die unverbindliche Überprüfung durch von diesen selbst besetzten Aufsichtsgremien, sind nicht ausreichend. Hier gibt es auf Ebene der Europäischen Union im Digital Services Act bereits Ansätze, die aber noch nicht weit genug gehen.

» *Empfehlung Kommunikation 2:* Transparenz über Moderations- und Kuratierungspraktiken: Anstelle allgemeiner Moderations- und Löschungsrichtlinien und wenig aussagekräftiger Zahlen über Löschungen muss für externe Kontrollen nachvollziehbar sein, wie, unter welchen Umständen und anhand welcher Kriterien solche Entscheidungen gefällt und umgesetzt werden und welche Rolle hierbei Algorithmen bzw. menschliche Moderierende übernommen haben. Darüber hinaus müssen auch die grundlegenden Funktionsweisen der Kuratierung von Inhalten sozialer Medien und Plattformen in dem Ausmaß offengelegt werden, das nötig ist, um systemische Verzerrungen und möglicherweise resultierende informationelle Dysfunktionen erkennen zu können. Die Berichtspflichten und Transparenzvorgaben im Medienstaatsvertrag, im Netzwerkdurchsetzungsgesetz und im Digital Services Act stellen dies noch nicht hinreichend sicher. Die datenschutzrechtlichen Auskunftspflichten gemäß Art. 12 ff. DSGVO sind zum Teil auf nationalstaatliche Ebene beschränkt worden und erfassen oftmals diese weiter gehenden Aspekte nicht.

» *Empfehlung Kommunikation 3:* Zugriff auf wissenschaftsrelevante Daten von Plattformen: Um die Wirkungsweisen von Plattformen und sozialen Medien, ihren Einfluss auf öffentliche Diskurse, aber auch weitere Themen von hoher gesellschaftlicher Relevanz zu untersuchen, sollte sichergestellt werden, dass unabhängigen Forschenden der Zugriff auf wissenschaftsrelevante Daten von Plattformen nicht

mit dem pauschalen Verweis auf Betriebs- oder Geschäftsgeheimnisse verweigert werden kann. Für den Zugang müssen sichere, datenschutzkonforme sowie forschungsethisch integre Wege gefunden werden. Netzwerkdurchsetzungsgesetz und Digital Services Act enthalten bereits Reglungen zum Datenzugang, die aber in ihrem Anwendungsbereich sehr begrenzt sind; auch der Data Act sieht vergleichbare Regelungen vor.

» *Empfehlung Kommunikation 4:* Berücksichtigung von Sicherheit, Datenschutz und Geheimhaltungsinteressen: Anforderungen an Offenlegungen und Datenzugang müssen kontextsensitiv spezifiziert werden, wobei Anforderungen an Sicherheit und Schutz vor Missbrauch, Datenschutz sowie dem Schutz von intellektuellem Eigentum und Geschäftsgeheimnissen angemessen Rechnung zu tragen ist. Je nach Kontext muss zwischen unterschiedlich klar definierten Zeitpunkten der Prüfung und Graden der Offenlegung unterschieden werden.

» *Empfehlung Kommunikation 5:* Personalisierte Werbung, Profiling und Microtargeting: Personalisierte Werbung ist das zentrale Geschäftsmodell sozialer Medien und Plattformen. Die Praktiken des Profilings und Microtargetings können jedoch problematische Auswirkungen auf öffentliche Kommunikation und Meinungsbildung entfalten, insbesondere im Kontext politischer Werbung. Um solche negativen Auswirkungen durch effektive Regelungen zu verhindern, ist es zunächst notwendig, die Bedingungen für eine Erforschung und Überprüfung der Zusammenhänge zwischen Geschäftsmodellen und Praktiken algorithmischer Kuratierung in ihren Wirkungsweisen und Effekten zu schaffen. Der auf Ebene der Europäischen Union diskutierte Vorschlag für eine Verordnung über die Transparenz und das Targeting

politischer Werbung adressiert diesen Bedarf. Hierbei zeigen sich allerdings auch die Herausforderungen, Regeln so zuzuschneiden, dass sie einerseits wirksam sind, andererseits aber die Freiheit der politischen Kommunikation nicht übermäßig beschränken.

» *Empfehlung Kommunikation 6:* Bessere Regulierung von Onlinemarketing und Datenhandel: Viele informationelle und kommunikative Dysfunktionen haben ihre Ursache im Onlinemarketing, welches das grundlegende Geschäftsmodell vieler sozialer Medien und Plattformen ist und auf der Sammlung, Analyse und dem Verkauf vielfältiger Daten über die Personen, die diese Angebote nutzen, beruht. Das Problem ist hierbei nicht die Werbefinanzierung per se, sondern der invasive Umgang mit diesen Daten. Hier gilt es einerseits, die Auswirkungen dieses Geschäftsmodells auf öffentliche Diskurse besser zu erforschen. Andererseits bedarf es besserer gesetzlicher Regelungen, um sowohl Individuen in ihren Grundrechten online effektiver zu schützen als auch negative systemische Effekte auf den öffentlichen Diskurs zu minimieren. In diese Richtung gehende Vorschläge hat der Deutsche Ethikrat 2017 unter dem Stichwort „Datensouveränität" in seiner Stellungnahme zu Big Data und Gesundheit vorgestellt. Europäische Regelungen wie der Digital Markets Act adressieren das Problem der Datenmacht großer Plattformen, aber – schon aus Gründen der Regelungskompetenz – nicht mit Blick auf die Folgen für den öffentlichen Diskurs.

» *Empfehlung Kommunikation 7:* Machtbeschränkung und Kontrolle: Unternehmen, die im Bereich der öffentlichen Vorstellung von Daten bzw. Tatsachen de facto monopolartige Machtmöglichkeiten haben, sind durch rechtliche Vorgaben und entsprechende Kontrolle auf Pluralismus, Minderheiten- und Diskriminierungsschutz zu verpflichten. Ein Teil der

Mitglieder des Deutschen Ethikrates ist der Auffassung, dass medienrechtliche Regelungen zur Sicherung von Pluralität, Neutralität und Objektivität generell auf Nachrichtenfunktionen von sozialen Medien und Plattformen ausgedehnt werden sollten, sofern sie denen traditioneller Medien ähneln.

» *Empfehlung Kommunikation 8:* Erweiterung der Nutzerautonomie: Plattformen und soziale Medien sollten ihre Inhalte auch ohne eine personalisierte Kuratierung verfügbar machen. Darüber hinaus sollten sie für die Kriterien, nach denen Inhalte auf Plattformen und in sozialen Medien algorithmisch ausgewählt und prioritär präsentiert werden, weitere Wahlmöglichkeiten anbieten. Dazu sollte auch die Möglichkeit gehören, bewusst Gegenpositionen angezeigt zu bekommen, die den bisher geäußerten eigenen Präferenzen zuwiderlaufen. Solche Wahlmöglichkeiten sollten gut sichtbar und leicht zugänglich sein.

» *Empfehlung Kommunikation 9:* Förderung kritischer Rezeption von Inhalten: Zur Eindämmung unreflektierter Verbreitung fragwürdiger Inhalte sollten diverse Hinweisfunktionen entwickelt und eingesetzt werden, die eine kritische Auseinandersetzung mit Material fördern, bevor man sich dafür entscheidet, es zu teilen oder öffentlich darauf zu reagieren. Dies könnten etwa Rückfragen sein, ob Texte gelesen und Videos geschaut wurden, bevor man sie teilt, oder Angaben zur Seriosität von Quellen.

» *Empfehlung Kommunikation 10:* Alternative Informations- und Kommunikationsinfrastruktur: Zu erwägen wäre, den privaten Social-Media-Angeboten im europäischen Rahmen eine digitale Kommunikationsinfrastruktur in öffentlich-rechtlicher Verantwortung zur Seite zu stellen, deren Betrieb sich nicht am Unternehmensinteresse eines möglichst langen Verweilens von Menschen auf der Plattform oder an

anderen kommerziellen Interessen orientiert. Damit sollte nicht etwa der öffentlich-rechtliche Rundfunk (TV und Radio) auf eine weitere digitale Plattform ausgedehnt, sondern eine digitale Infrastruktur bereitgestellt werden, die eine Alternative zu den kommerzbetriebenen, stark oligopolartigen Angeboten bietet. Um eine hinreichende Staatsferne zu garantieren, könnte auch an eine Trägerschaft in Gestalt einer öffentlichen Stiftung gedacht werden.

Öffentliche Verwaltung

104) Für viele Menschen und Organisationen stellt die öffentliche Verwaltung, so etwa im Finanz-, Steuer-, Melde- und Sozialwesen und in der Straffälligen- und Jugendgerichtshilfe, die unmittelbar erfahrbare Staatsgewalt dar. Funktionierende, transparente, als legitim anerkannte und bürgernahe Verwaltung ist für ein funktionierendes Gemeinwesen und die Akzeptanz von Demokratie und Staat wesentlich. Mit Digitalisierungsstrategien in diesem Bereich verbinden sich Hoffnungen auf eine Rationalisierung und Beschleunigung staatlichen Verwaltungshandelns, eine effektivere und kohärentere Datennutzung sowie eine Ausweitung der Einbeziehung wissenschaftlichen und bürgerschaftlichen Sachverstandes. Dem steht die dystopische Schreckensvision einer sogenannten „Algokratie" gegenüber, in der autonome Softwaresysteme die staatliche Herrschaft über Menschen ausüben.

105) Vielfach und zunehmend werden in der öffentlichen Verwaltung automatisierte Entscheidungssysteme (Automated/Algorithmic Decision Making Systems, ADM-Systeme) eingesetzt, etwa zur Bewertung von Arbeitsmarktchancen, bei der Prüfung und Vergabe von Sozialleistungen oder für Vorhersagen im Bereich der Polizeiarbeit. Von besonderem Interesse ist hier, inwieweit der Einsatz von KI-Systemen menschliche Handlungsfähigkeiten und Autorschaft beeinflusst. Angesichts

der häufig beobachteten Tendenz, sich maschinellen Empfehlungen vorbehaltslos anzuschließen (Automation Bias), kann bereits die Nutzung von Software zur Entscheidungsunterstützung in der Verwaltung weitreichende Wirkung entfalten.

106) Andere Fragen betreffen vor allem Aspekte von Gerechtigkeit, beispielsweise wenn es darum geht, ob und in welchem Umfang die verwendeten Systeme Diagnosen und Prognosen tatsächlich verbessern, ob die Genauigkeit für verschiedene Anwendungskontexte oder Personengruppen gleich ist oder ob es systematische Verzerrungen oder Diskriminierungen gibt (Algorithmic Bias). Ebenso können datenbasierte Systeme jedoch historische Ungerechtigkeiten oder menschliche Vorurteile aufdecken und sie damit für Gegenmaßnahmen zugänglich machen.

107) Eine grundsätzliche Grenze für die Anwendung von automatisierten Entscheidungssystemen liegt in nicht eliminierbaren normativen Ziel- oder Regelkonflikten im deutschen, deontologisch verfassten Rechtssystem, in dem die Folgenabwägung nie allein das Rechtmäßige bestimmt, sondern unbedingte Ansprüche auf Schutz der Person zu wahren sind und der Algorithmisierung ethischer und rechtlicher Entscheidungsprozesse Grenzen setzen.

108) Das Sozialwesen ist ein Bereich der Verwaltung, in dem Entscheidungen mit weitreichenden Folgen für die Betroffenen fallen, etwa über die Gewährung von Hilfen, bei Maßnahmen im Kontext einer Kindeswohlgefährdung oder bei der Abschätzung von Gefährdungspotenzialen von Straffälligen in der Bewährungshilfe. Algorithmenbasierte Entscheidungshilfen kommen hier zunehmend zum Einsatz und können professionelle Handlungskompetenz erweitern, wenn sie Fachkräften helfen, ihre sonst oft intuitiven Einschätzungen auf eine solidere Datengrundlage zu stellen, bei Bedarf zu korrigieren und

Entscheidungen so evidenzbasiert zu standardisieren. Dies ist besonders wichtig bei der Abschätzung von Gefährdungspotenzialen, beispielsweise bei Verdacht auf Kindeswohlgefährdung oder in der Bewährungshilfe.

109) Menschliche Autorschaft kann unter Zuhilfenahme sachdienlicher Ergebnisse von KI-Algorithmen allerdings auch vermindert werden. Auf der professionellen Seite kann dies beispielsweise dann der Fall sein, wenn es zu einer ungeprüften Übernahme algorithmisch vorgeschlagener Ergebnisse kommt (Automation Bias). Auch für die von den Entscheidungen betroffenen Personen sind negative Effekte möglich, etwa wenn ihnen aufgrund von durch Verzerrungen geprägten algorithmisch unterstützten Entscheidungen Handlungs- oder Entwicklungsmöglichkeiten ungerechtfertigterweise genommen werden.

110) Gerade bei der Erfassung von Hilfebedarfen birgt der Einsatz algorithmischer Systeme zudem das Risiko der Entkopplung aus einer dialogischen Beziehungsarbeit, die entscheidend für die Erfahrung von Selbstwirksamkeit Betroffener sein kann. Wird diese persönliche Ebene bei der Ermittlung des individuellen Hilfebedarfs im Zuge einer algorithmenbasierten Informatisierung des Sozialwesens vernachlässigt, können positive Effekte selbst materieller Hilfeleistungen schnell verpuffen und damit kaum nachhaltig wirken. Das österreichische Arbeitsmarktchancen-Assistenzsystem beispielsweise, das für Arbeitssuchende Erfolgsprognosen für eine Wiedereingliederung in den Arbeitsmarkt berechnet, ist für seine Ausrichtung an den Werten, Normen und Zielen einer restriktiven Fiskalpolitik kritisiert worden, die den Zielen eines personenorientierten Hilfesystems, welches individuelle Hilfebedarfe betroffener Personen fokussieren muss, diametral zuwiderläuft.

111) Ein anderer Bereich, in dem algorithmenbasierte Risikoanalysen zunehmend zum Einsatz kommen, ist die

Kriminalitätsbekämpfung. Im Predictive Policing unterstützen entsprechende Anwendungen präventive Polizeiarbeit mittels Prognosen künftiger Straftaten, straffälliger Personen und Tatorte, um Verbrechen zu verhindern. Vor allem personenbezogene Verfahren werden in diesem Zusammenhang kontrovers diskutiert. Einerseits geht damit die Hoffnung auf eine bessere polizeiliche Arbeit und einen besseren Schutz möglicher Opfer einher. Andererseits können Fehler und Verzerrungen in algorithmenbasierter Verbrechensbekämpfung mit besonders folgenschweren Konsequenzen für ungerechtfertigt klassifizierte Personen verbunden sein und zudem in der Software systembedingt besondere Breitenwirkung entfalten.

112) Ein weiteres Problem ist der Schutz der Privatsphäre im Kontext von Predictive Policing. Die für die Polizeiarbeit herangezogenen Daten sind in aller Regel besonders sensibel. Insbesondere bei sogenannten Chatkontrollen zur Prävention und Bekämpfung des sexuellen Missbrauchs von Kindern, zu denen die Europäische Kommission im Mai 2022 einen Verordnungsvorschlag vorgelegt hat, wird hinterfragt, ob eine anlasslose und flächendeckende Überwachung privater Kommunikation gerechtfertigt werden kann oder einen unverhältnismäßig intensiven Eingriff in die Grundrechte darstellt.

113) Nicht zuletzt wird die Sorge geäußert, dass mit algorithmengesteuerter Polizeiarbeit das Risiko der Verfestigung eines mechanischen Menschenbildes einhergehen könne, welches den einzelnen Menschen verobjektiviere, seine Individualität auf eine datengetriebene Klassifikation reduziere, jedoch die gesamtgesellschaftlichen Ursachen von Kriminalität unberücksichtigt lasse.

114) In der Zusammenschau führen automatisierte Entscheidungsverfahren in der öffentlichen Verwaltung zu neuen Möglichkeiten und Herausforderungen, die erheblich weiter reichende ethische und demokratietheoretische

Fragen aufwerfen, etwa in Bezug auf Nachvollziehbarkeit, Erklärbarkeit und Vertrauenswürdigkeit im Verwaltungshandeln, aber auch in Bezug auf Sorgen um Diskriminierung und Technokratie, in der menschliche Kommunikation und Abwägung hinter anonymen Datenmengen und standardisierten Bedienoberflächen verschwinden.

115) Insofern der Rückgriff auf große Datenmengen und ihre zielgerichtete Auswertung bessere Entscheidungsgrundlagen schaffen, kann der Einsatz algorithmischer Systeme zu diesem Zweck menschliche Autorschaft unterstützen und ist in ethischer Hinsicht grundsätzlich zu begrüßen. Eine unkritische Übernahme von Systemempfehlungen droht menschliche Autorschaft allerdings zu vermindern, bis im ungünstigen Fall nur noch ein automatisiertes Geschehen verbleibt, in dem technische Systeme für Betroffene weitreichende, teils existenzielle Festlegungen treffen und systemimmanente Fehler oder Verzerrungen möglicherweise nicht mehr erkannt werden.

116) Beim Einsatz von KI in der öffentlichen Verwaltung muss daher kontextbezogen im Detail eingeschätzt und abgewogen werden, welche Auswirkungen eine entsprechende Maßnahme auf die Autorschaft unterschiedlichster Beteiligter und Betroffener hat, welche Konflikte auftreten und wie mit ihnen umgegangen werden kann oder soll. Hierzu legt der Deutsche Ethikrat neun Empfehlungen vor:

» *Empfehlung Verwaltung 1:* Die mit automatisierten Entscheidungshilfen (ADM-Systeme) einhergehende verstärkte Standardisierung und pauschale Kategorisierung von Einzelfällen müssen umso stärker hinterfragt und um spezifisch einzelfallbezogene Erwägungen ergänzt werden, je intensiver die betroffene Entscheidung individuelle Rechtspositionen berührt.

- *Empfehlung Verwaltung 2:* Es müssen geeignete technische und organisatorische Instrumente zur Vorkehrung gegen die manifeste Gefahr eines Automation Bias bereitgestellt werden, die es den Fachkräften erschweren, selbst bei einer Letztentscheidungskompetenz der algorithmischen Entscheidungsempfehlung unbesehen zu folgen. Es ist zu prüfen, ob eine Umkehrung der Begründungspflicht (nicht eine Abweichung, sondern ein Befolgen ist zu rechtfertigen) hier eine geeignete Vorkehrung sein kann.
- *Empfehlung Verwaltung 3:* Aufgrund ihrer Grundrechtsbindung sind an staatliche Einrichtungen bei der Entwicklung und Nutzung algorithmischer Systeme hohe Anforderungen in Bezug auf Transparenz und Nachvollziehbarkeit zu stellen, um den Schutz vor Diskriminierung zu gewährleisten sowie Begründungspflichten erfüllen zu können.
- *Empfehlung Verwaltung 4:* Für Softwaresysteme in der öffentlichen Verwaltung müssen Qualitätskriterien verbindlich und transparent festgelegt werden (z. B. in Bezug auf Genauigkeit, Fehlervermeidung und Unverzerrtheit). Ebenso bedarf es einer Dokumentation der jeweils eingesetzten Methoden. Diesbezüglich sollten auch aktuelle Beschaffungspraktiken, in deren Verlauf staatliche Behörden Softwarelösungen kaufen, einer kritischen Prüfung unterzogen werden.
- *Empfehlung Verwaltung 5:* Überall dort, wo algorithmische Systeme Einsatz in der öffentlichen Verwaltung finden, gilt es, Sorge zu tragen, dass die Personen, die diese Systeme anwenden, über die erforderlichen Kompetenzen im Umgang damit verfügen. Dazu gehört neben der Kenntnis der Verwendungsweisen auch das Wissen um die Limitationen und möglichen Verzerrungen, um Systeme angemessen einsetzen zu können.

- *Empfehlung Verwaltung 6:* Die Einsichts- und Einspruchsrechte Betroffener müssen auch beim Einsatz algorithmischer Systeme effektiv gewährleistet werden. Dazu bedarf es gegebenenfalls weiterer wirksamer Verfahren und Institutionen.
- *Empfehlung Verwaltung 7:* In Öffentlichkeit, Politik und Verwaltung sollte eine Sensibilisierung gegenüber möglichen Gefahren von Automatisierungssystemen, wie etwa Verletzungen der Privatsphäre oder Formen systematisierter Diskriminierung, erfolgen. Dazu gehört eine öffentliche Debatte darüber, ob es in bestimmten Kontexten überhaupt einer technischen Lösung bedarf.
- *Empfehlung Verwaltung 8:* Im Bereich des Sozialwesens ist sicherzustellen, dass ADM-Systeme elementare fachliche Standards von sozialprofessionellen Interaktionen (z. B. gemeinsame Sozialdiagnose oder Hilfeplanung als *Teil* therapeutischer bzw. unterstützender Hilfeleistung) nicht unterlaufen oder verdrängen. Dies beinhaltet insbesondere Maßnahmen, die Vergröberungen individueller Fallkonstellationen und -prognosen durch die ADM-induzierte grobklassifizierende Einteilung von Fall- und/oder Leistungsberechtigtengruppen verhindern. Dabei ist Sorge zu tragen, dass die Feststellung individueller Hilfebedarfe nicht erschwert wird und es zu keiner schleichenden Aushöhlung der sozialrechtlich gebotenen Identifizierung individueller Hilfebedarfe zugunsten einseitiger externer Interessen an Gefahrenminimierung oder Kostenersparnis kommt.
- *Empfehlung Verwaltung 9:* Die Arbeit von Gefahrenabwehrbehörden einschließlich der Polizei betrifft besonders grundrechtssensible Bereiche. Dies wirkt sich auf die Reichweite eines zulässigen Einsatzes von algorithmischen Systemen in der prädiktiven Polizeiarbeit aus. Risiken wie Verletzungen der Privatsphäre

oder potenziell unzulässige Diskriminierungen der von dem Einsatz betroffenen Personen müssen mit Chancen auf erhebliche Verbesserungen der staatlichen Gefahrenabwehr sorgfältig abgewogen und in ein angemessenes Verhältnis gebracht werden. Hierfür erforderliche gesellschaftliche Aushandlungsprozesse sollten umfangreich geführt werden. Dabei ist der diffizilen Bestimmung des Verhältnisses von Freiheit und Sicherheit Rechnung zu tragen. Jegliche Gesetzesübertretung zu verhindern, wäre mit rechtsstaatlichen Mitteln nicht möglich.

TEIL III: QUERSCHNITTSTHEMEN UND ÜBERGREIFENDE EMPFEHLUNGEN

Zusammenfassung der bisherigen Analyse

117) Der Begriff der Künstlichen Intelligenz hat in der öffentlichen Debatte zunehmend an Aufmerksamkeit gewonnen und wird teils mit überzogenen Hoffnungen, teils aber auch mit fehlgeleiteten Befürchtungen verknüpft. Der Deutsche Ethikrat geht von einem normativ grundlegenden Unterschied zwischen Mensch und Maschine aus. Softwaresysteme verfügen weder über theoretische noch über praktische Vernunft. Sie handeln oder entscheiden nicht selbst und können keine Verantwortung übernehmen. Sie sind kein personales Gegenüber, auch dann nicht, wenn sie Anteilnahme, Kooperationsbereitschaft oder Einsichtsfähigkeit simulieren.

118) Menschliche Vernunft ist immer zugleich eingebunden in die konkrete soziale Mit- und Umwelt. Nur so ist zu erklären, dass sie handlungswirksam wird. Vernünftig handelt der einzelne Mensch als Teil einer sozialen Mitwelt und einer kulturellen Umgebung. Schon deshalb kann den in dieser Stellungnahme besprochenen

Softwaresystemen weder theoretische noch praktische Vernunft zugeschrieben werden.

119) Menschen entwickeln digitale Technik und nutzen sie als Mittel zu menschlichen Zwecken. Jedoch wirken diese Technologien zurück auf menschliche Handlungsmöglichkeiten. Sie können einerseits neue Optionen eröffnen, aber andererseits auch Anpassungen erforderlich machen, die nicht wünschenswert sind. Auch wenn Maschinen also nicht selbst handeln, so verändern sie die Handlungsfähigkeit von Menschen tiefgreifend und können Handlungsmöglichkeiten erheblich erweitern oder vermindern.

120) Ziel der Delegation menschlicher Tätigkeiten an Maschinen sollte prinzipiell die Erweiterung menschlicher Handlungsfähigkeit und Autorschaft sein. Ihre Verminderung sowie eine Diffusion oder Evasion von Verantwortung gilt es hingegen zu verhindern. Dafür muss die Übertragung menschlicher Tätigkeiten auf KI-Systeme gegenüber den Betroffenen hinreichend transparent erfolgen, sodass wichtige Entscheidungselemente, -parameter oder -bedingungen nachvollziehbar bleiben.

121) Um über Wert und Nutzen der Delegation vormals menschlichen Handelns an Maschinen ethisch zu befinden, bedarf es daher immer eines kontextspezifischen Blicks, der die Perspektiven unterschiedlicher Beteiligter und Betroffener ebenso berücksichtigt wie die langfristigen Auswirkungen solcher Übertragungen. Die Herausforderungen stecken also wie so oft im Detail, genauer: in den Details der Technik, der Einsatzkontexte sowie der institutionellen und soziotechnischen Umgebung.

122) Um diesen kontextspezifischen Blick zu ermöglichen, hat sich der Deutsche Ethikrat in dieser Stellungnahme exemplarisch mit Anwendungen in der Medizin, der schulischen Bildung, der öffentlichen Kommunikation sowie der Verwaltung beschäftigt. Es wurden bewusst Sektoren ausgewählt, in denen die Durchdringung

durch KI-basierte Technologien sehr unterschiedlich ausfällt und sich jeweils unterschiedliche Ausmaße des Ersetzens vormals menschlicher Handlungen durch KI veranschaulichen lassen. In allen vier Sektoren sind Einsatzszenarien durch teils erhebliche Beziehungs- und Machtasymmetrien gekennzeichnet, was einen verantwortungsvollen Einsatz von KI und die Berücksichtigung der Interessen und des Wohls insbesondere vulnerabler Personengruppen umso wichtiger macht. Diese Unterschiedlichkeit der Art und Weise des KI-Einsatzes sowie des Ausmaßes der Delegation an Maschinen in den Blick zu nehmen, erlaubt es nuancierte ethische Betrachtungen anzustellen.

Entfaltung von Querschnittsthemen und Empfehlungen

123) Die Darstellung der soziotechnischen Entwicklungen und deren ethische Analyse in den vier Anwendungsbereichen zeigen, dass es eine Reihe von Querschnittsthemen und -herausforderungen gibt, die sich durch alle vier Bereiche ziehen, wenn auch teils in unterschiedlicher Weise und Ausprägung. Um im Hinblick auf die Erweiterung menschlicher Handlungsfähigkeit und Autorschaft zukünftig einen guten gesellschaftlichen Umgang mit KI zu gewährleisten, müssen solche Querschnittsfragen nicht nur innerhalb einzelner Bereiche angegangen werden, sondern darüber hinaus auch in vernetzten, bereichsübergreifenden Ansätzen.

124) Solches gleichermaßen horizontale wie vertikale gestaltende Denken stellt eine Herausforderung insbesondere für die Politikgestaltung und etwaige zukünftige Regulierung dar. Die Darstellung der Querschnittsthemen in dieser Stellungnahme, die für jedes Thema in einer Empfehlung münden, soll daher als Anregung für eine breitere Debatte dienen, wie für zukünftige Politik- und Technikgestaltung gleichzeitig und im Zusammenspiel mit

sektoralen Aspekten immer auch übergreifende Fragen in den Blick genommen werden können und müssen.

125) Im *ersten Querschnittsthema* geht es noch einmal um das in dieser Stellungnahme zentrale Konzept der Erweiterung und Verminderung menschlicher Handlungsmöglichkeiten. Zwar besteht eine sektorübergreifende Gemeinsamkeit hinsichtlich der angestrebten Erweiterung menschlicher Handlungspotenziale darin, dass die komplette Ersetzung menschlicher Akteure durch KI-Systeme sich überall dort verbietet, wo die konkrete zwischenmenschliche Begegnung eine notwendige Voraussetzung für die Erreichung der jeweiligen Handlungsziele darstellt. Darüber hinaus besteht jedoch die Notwendigkeit, die Unterschiede beim KI-Einsatz in den einzelnen Handlungsbereichen sorgfältig zu beachten.

» *Empfehlung Querschnittsthema 1:* Da die Vor- und Nachteile von KI-Anwendungen für verschiedene Personengruppen sowie die Gefahr des Verlustes bestimmter Kompetenzen bei den Personen, die solche Systeme anwenden, erheblich variieren, bedarf es sowohl einer differenzierten Planung des KI-Einsatzes in unterschiedlichen Handlungsfeldern, welche die jeweiligen Zielsetzungen und Verantwortlichkeiten präzise benennt, als auch einer zeitnahen Evaluation der tatsächlichen Folgen eines solchen Einsatzes, um die Systeme besser an die spezifischen Handlungskontexte anzupassen und sie fortlaufend zu verbessern.

126) Das *zweite Querschnittsthema* behandelt Wissenserzeugung durch KI und den Umgang mit KI-gestützten Voraussagen. Zentral ist dabei die Prämisse, dass Korrelationen und Datenmuster nicht mit Erklärungen und Begründungen von Ursachen von Ereignissen gleichzusetzen sind, sondern auch qualitativ evaluiert und normativ beurteilt werden müssen. Bei probabilistischen

Methoden bleiben immer Restunsicherheiten, über deren Akzeptabilität zu entscheiden ist. Ethisch positiv zu werten ist, dass durch KI-Einsatz in allen vier hier betrachteten Anwendungsbereichen erhebliche funktionale Verbesserungen erreicht wurden und weiterhin erwartbar sind. Es wird jedoch eine grundsätzlich normativ problematische Schwelle überschritten, wenn funktionale Verbesserungen (eventuell sogar unbemerkt) in eine Ersetzung moralischer Kompetenz und damit verbundener Verantwortung hinübergleiten.

» *Empfehlung Querschnittsthema 2:* Der Einsatz KI-gestützter digitaler Techniken ist im Sinne der Entscheidungsunterstützung und nicht der Entscheidungsersetzung zu gestalten, um Diffusion von Verantwortung zu verhindern. Er darf nicht zulasten effektiver Kontrolloptionen gehen. Den von algorithmisch gestützten Entscheidungen Betroffenen ist insbesondere in Bereichen mit hoher Eingriffstiefe die Möglichkeit des Zugangs zu den Entscheidungsgrundlagen zu gewähren. Das setzt voraus, dass am Ende der technischen Prozeduren entscheidungsbefugte Personen sichtbar bleiben, die in der Lage und verpflichtet sind, Verantwortung zu übernehmen.

127) Das *dritte Querschnittsthema* betrachtet die Gefährdung des Individuums durch statistische Stratifizierung. Grundlage vieler KI-Anwendungen sind Korrelationen, die bei der Analyse großer Datenmengen entdeckt werden und anhand derer man Einzelpersonen Kohorten mit bestimmten Merkmalskombinationen zuordnen kann. Die Bildung solcher Kohorten und die auf ihrer Basis durch Algorithmen produzierten Voraussagen können die Qualität und Effektivität einer Anwendung insgesamt verbessern. Sie können aber auch Probleme für Individuen bedeuten, welche von solchen kollektiven

Schlüssen betroffen sind – insbesondere dann, wenn die statistisch getroffene Diagnose oder Prognose in ihrem Fall nicht zutrifft.

» *Empfehlung Querschnittsthema 3:* Neben einer Analyse der konkreten und naheliegenden Probleme datenbasierter Software, beispielsweise in Bezug auf den Schutz der Privatsphäre oder die Verhinderung von Diskriminierung, gilt es, auch die langfristigen Auswirkungen dieser statistischen Präkonfiguration von Individuen sowie deren Rückwirkung – im Sinne einer Erweiterung oder Verminderung der Handlungsmöglichkeiten – auf Individuen wie Kollektive für alle Sektoren sorgfältig zu beleuchten. Darüber hinaus gilt, dass Einzelfallbeurteilungen grundsätzlich wichtig bleiben. KI-basierte Beurteilungen und Vorhersagen können unter günstigen Bedingungen ein Hilfsmittel sein, aber kein geeignetes Instrument der *definitiven* Lagebeurteilung und Entscheidung. Pragmatische und heuristische Faktoren wie die Prüfung der Kohärenz mit anderen Evidenzquellen oder Erfolgseinschätzungen spielen eine nicht zu vernachlässigende Rolle.

128) Im *vierten Querschnittsthema* geht es um die Auswirkungen von KI auf menschliche Kompetenzen und Fertigkeiten. Deren Erwerb und Erhalt kann durch die Delegation menschlicher Tätigkeiten an Maschinen gefährdet werden. Weil die Nutzung von KI-Anwendungen (wie auch bei anderen Technologien) dazu führen kann, dass menschliche Fähigkeiten nachlassen bzw. ganz verkümmern, können Abhängigkeiten von diesen Technologien entstehen. Handelt es sich dabei um gesellschaftlich besonders bedeutsame oder kritische Einsatzbereiche, ist ein Verlust von menschlichen Kompetenzen und Fertigkeiten ein ernstzunehmendes Risiko.

» *Empfehlung Querschnittsthema 4:* Ob und inwiefern beim Einsatz von KI-Anwendungen Verluste menschlicher Kompetenz auftreten, die als unerwünscht eingestuft werden, muss sorgfältig beobachtet werden. Bei der Entwicklung und dem Einsatz neuer Technologien sind solch unerwünschte Kompetenzverluste durch eine sinnvolle Gestaltung des Zusammenspiels von Mensch und Technik, durch angemessene institutionelle und organisatorische Rahmenbedingungen sowie durch gezielte Gegenmaßnahmen wie etwa spezifische Trainingsprogramme zu minimieren bzw. zu kompensieren. Kompetenzverluste können sowohl individueller als auch kollektiver Natur sein. So gilt es zu verhindern, dass die Delegation von Aufgaben an Technologien dazu führt, dass Gesellschaften übermäßig anfällig werden, wenn diese Technologien (zeitweise) ausfallen. Jenseits dieser systemischen Aspekte müssen negative Auswirkungen solcher Delegation auf die individuelle Autonomie oder Selbstwahrnehmung mitigiert werden.

129) Das *fünfte Querschnittsthema* befasst sich mit dem Schutz von Privatsphäre und Autonomie versus Gefahren durch Überwachung und Chilling-Effekte. Die im Rahmen vieler KI-Anwendungen notwendige Erfassung großer Mengen an personenbezogenen Daten sowie die Möglichkeit, auf ihrer Basis sensible Prognosen zu erstellen, beeinträchtigen nicht nur die Privatsphäre der Personen, von denen diese Daten stammen, sondern machen sie auch vulnerabel gegenüber möglichen Benachteiligungen oder Manipulation, welche aus der Verarbeitung der Daten resultieren können. Chilling-Effekte beschreiben in diesem Kontext Rückwirkungen auf das Verhalten von Menschen, die Sorge haben, dass ihr Verhalten beobachtet, aufgezeichnet oder ausgewertet wird.

» *Empfehlung Querschnittsthema 5:* Die beschriebenen Phänomene sollten in ihrer Entstehung, Ausprägung und Entwicklung umfassend empirisch untersucht werden. Um sowohl dem Problem der Überwachung sowie den parallelen Gefahren durch etwaige Chilling-Effekte Rechnung zu tragen, müssen angemessene und effektive rechtliche und technische (z. B. Privacy by Design) Vorkehrungen getroffen werden, die dem übermäßigen Tracking von Onlineverhalten und dem Handel mit personenbeziehbaren Daten Einhalt gebieten. Die Interessen der Datensubjekte müssen hierbei im Mittelpunkt stehen. Insbesondere ist dabei auf besonders vulnerable Gruppen zu achten, da viele der Einsatzkontexte zudem von asymmetrischen Machtverhältnissen gekennzeichnet sind. Es muss Sorge getragen werden, dass die Erweiterung der Handlungsmöglichkeiten einiger nicht zulasten der Verminderung der Handlungsmöglichkeiten anderer, insbesondere benachteiligter Gruppen stattfindet.

130) Das *sechste Querschnittsthema* greift Konzepte von Datensouveränität und gemeinwohlorientierter Datennutzung auf, die der Deutsche Ethikrat bereits 2017 in seiner Stellungnahme zu Big Data und Gesundheit entwickelt hat. Dabei geht es um die Suche nach Lösungen, wie im Kontext von KI-Anwendungen Daten sinnvoll für verschiedene wichtige Zwecke genutzt werden können, ohne zugleich den Schutz der Privatsphäre von Datensubjekten unzulässig zu beeinträchtigen. Hier stellt sich die Frage, ob das derzeitige Datenschutzrecht bzw. die herrschende Datenschutzpraxis diesen beiden Zielen gerecht wird. Während in manchen Handlungsfeldern berechtigte Sorgen vor unbemerkten und weitreichenden Verletzungen von Privatsphäre und informationeller Selbstbestimmung herrschen, werden in anderen Kontexten durch strenge Auslegungen von Datenschutzregeln wichtige

soziale Güter, etwa mit Blick auf Patientenversorgung und wissenschaftlichen Erkenntnisgewinn, aber auch die kommunale Daseinsvorsorge nicht oder nur sehr schwer erreicht.

» *Empfehlung Querschnittsthema 6:* Mit Blick auf KI-Anwendungen müssen neue Wege gefunden werden, um innerhalb der jeweiligen Kontexte und bezüglich der jeweils spezifischen Herausforderungen und Nutzenpotenziale die gemeinwohlorientierte Daten(sekundär)nutzung zu vereinfachen bzw. zu ermöglichen und damit die Handlungsoptionen auf diesem Gebiet zu erweitern. Zugleich ist es essenziell, einen Bewusstseinswandel sowohl in der Öffentlichkeit als auch bei den praktisch tätigen Personen, die Datennutzung gestalten, herbeizuführen – weg von einer vornehmlich individualistisch geprägten und damit verkürzten Perspektive hin zu einer Haltung, die auch systematische und gemeinwohlbasierte Überlegungen mit einbezieht und in einen Ausgleich bringt. Eine solche Haltung ist auch für die zukünftige Politikgestaltung und Regulierung deutlich stärker als bisher zugrunde zu legen. Nur so kann es gelingen, neben den Risiken, die sich aus breiterer KI-Anwendung ohne Zweifel ergeben, zugleich die wichtigen Chancen einer verantwortlichen Nutzung nicht aus dem Blick zu verlieren.

131) Das *siebte Querschnittsthema* betrachtet kritische Infrastrukturen, Abhängigkeiten und Resilienz. Im Zuge der Digitalisierung werden Infrastrukturen wie beispielsweise Stromnetze zunehmend digital überwacht und über das Internet gesteuert. Gleichzeitig werden digitale Technologien selbst zu Infrastrukturen. Am Vorhandensein und Funktionieren von Infrastrukturen richten Menschen ihr Handeln aus, und im Zuge dieser sozialen

Aneignung entstehen Abhängigkeiten, die menschliche Autonomie gefährden können. Wenn KI-gestützte Systeme zusehends in die Steuerung von Infrastrukturen integriert werden, kommt hinzu, dass solche Systeme nicht vollständig transparent und nachvollziehbar sind, und durch die fortwährende Komplexitätssteigerung von Infrastruktursystemen und ihrer Steuerung steigt die gesellschaftliche und institutionelle Vulnerabilität weiter an.

» *Empfehlung Querschnittsthema 7:* Um die Autorschaft menschlicher Akteure und deren Handlungsmöglichkeiten zu erweitern, müssen die Resilienz soziotechnischer Infrastrukturen gestärkt und die Abhängigkeit von individuellen Akteuren und Systemen minimiert werden. Dies umfasst zunächst die Notwendigkeit, die infrastrukturelle Bedeutung digitaler Technologien anzuerkennen und infolgedessen dem Schutz und der Resilienz kritischer digitaler Infrastrukturen mehr Aufmerksamkeit zuteilwerden zu lassen, auch im politischen Handeln. In allen Sektoren gilt es, einseitige Abhängigkeiten zu vermeiden, welche im Krisenfalle verletzlich und angreifbar machen. Für Nutzerinnen und Nutzer erfordert eine Verringerung der Abhängigkeit die Möglichkeit, zwischen Alternativen zu wählen, ohne große Teile der Funktionalität einzubüßen. Dies umfasst die Notwendigkeit von Interoperabilität, um einfach zwischen Systemen wechseln zu können. Hierfür ist auch der Auf- und Ausbau alternativer Infrastrukturen von besonderer Bedeutung. Im Kontext der öffentlichen Meinungsbildung erscheint die Etablierung unabhängiger, öffentlicher digital-kommunikativer Plattformen dringend geboten. Aber auch in anderen Sektoren wie der Verwaltung, der Bildung oder der Medizin vermindert eine zu große Abhängigkeit von wenigen Systemen oder

Akteuren potenziell die individuelle wie kollektive Handlungsfähigkeit.

132) Das *achte Querschnittsthema* dreht sich um Pfadabhängigkeiten, Zweitverwertung und Missbrauchsgefahren. Pfadabhängigkeiten entstehen, wenn Entscheidungen, die zu Beginn einer bestimmten Entwicklung getroffen wurden, noch lange nachwirken und teils schwer wieder aufzuheben sind, auch wenn sich der Kontext der Nutzung möglicherweise geändert hat. Sind Technologien einmal eingeführt, dürfte zudem eine Tendenz auszumachen sein, deren Möglichkeiten voll auszuschöpfen – auch über das ursprüngliche Anwendungsfeld hinaus. Solche Zweitverwertungen sind nicht prinzipiell problematisch, doch sobald eine Technologie etabliert ist, kann es schwer sein, weitere, auch missbräuchliche Nutzungsszenarien auszuschließen. Gerade digitale Technologien und insbesondere Grundlagentechnologien wie das maschinelle Lernen eröffnen oft sehr mannigfaltige Nutzungsmöglichkeiten, in denen die Frage der Abgrenzung von Ge- und Missbrauch zunehmend schwieriger wird.

» *Empfehlung Querschnittsthema 8:* Bei Technologien mit großen Auswirkungen oder hohem Verbreitungsgrad und vor allem dort, wo sich eine Nutzung von Technologien kaum oder gar nicht vermeiden lässt, müssen bereits zu Beginn der Entwicklungsplanung mögliche Langzeitfolgen wie Pfadabhängigkeiten im Allgemeinen sowie Dual-Use-Potenziale im Speziellen regelhaft und explizit mitgedacht und antizipiert werden. Dies gilt in besonderem Maße in der Anwendungsplanung. Dabei sind neben direkten, sektorspezifischen Schadenspotenzialen auch etwaige – natürlich deutlich schwieriger fass- und antizipierbare – sektorübergreifende Effekte zu bedenken. Hohe Standards für die Sicherheit und den Schutz der

Privatsphäre (Security by Design, Privacy by Design) können ebenfalls dazu beitragen, spätere missbräuchliche Anwendungen einzuhegen bzw. möglichst zu verhindern. Bei besonders invasiven Technologien beispielsweise in der öffentlichen Verwaltung, die Bürgerinnen und Bürger gegebenenfalls verpflichtend nutzen müssen, sind besonders hohe Standards einzuhalten. Um dies sicherzustellen und überprüfen zu können, sind gegebenenfalls Open-Source-Ansätze angezeigt.

133) Im *neunten Querschnittsthema* geht es um Bias und Diskriminierung. Datenbasierte KI-Systeme lernen auf Basis vorhandener Daten. Resultierende Prognosen und Empfehlungen schreiben somit die Vergangenheit in die Zukunft fort, wodurch Stereotypen, aber auch bestehende gesellschaftliche Ungleichheiten und Ungerechtigkeiten durch den Einbau in scheinbar neutrale Technologien reproduziert und sogar verstärkt werden können. Oft liegt bei der Entwicklung von KI-Systemen keine unmittelbare Diskriminierungsabsicht vor, sondern diskriminierende Effekte entstehen aus gesellschaftlichen Realitäten oder Stereotypen in Kombination mit technisch-methodischen Entscheidungen. Es ist allerdings zumindest denkbar, dass auch explizite Diskriminierungsabsichten in komplexen Systemen versteckt werden könnten.

» *Empfehlung Querschnittsthema 9:* Zum Schutz vor Diskriminierung in Anbetracht der zuvor dargelegten Herausforderungen bedarf es *angemessener Aufsicht und Kontrolle* von KI-Systemen. Besonders in sensiblen Bereichen erfordert dies den Auf- oder Ausbau gut ausgestatteter Institutionen. Hier gilt: je größer die Eingriffstiefe und je unumgänglicher die Systeme, desto höher die Anforderungen an Diskriminierungsminimierung. Auch bereits bei der Entwicklung von

Technologien gilt es, Diskriminierung zu minimieren bzw. Fairness, Transparenz und Nachvollziehbarkeit herzustellen. Dies sollte sowohl durch Anreize – etwa Forschungsförderung – als auch durch entsprechende gesetzliche Anforderungen befördert werden, etwa hinsichtlich der Offenlegung, welche Maßnahmen zur Diskriminierungsminimierung bei der Softwareentwicklung ergriffen wurden. Allerdings haben technische wie regulatorische Maßnahmen zur Minimierung von Diskriminierung ihre Grenzen, unter anderem weil unterschiedliche Fairnessziele technisch nicht gleichzeitig erfüllt werden können. Es müssen also zugleich ethische und politische Entscheidungen getroffen werden, welche Kriterien für Gerechtigkeit in welchem Kontext zum Tragen kommen sollen. Diese Entscheidungen dürfen nicht den Personen, die Software entwickeln, und anderen direkt Beteiligten überlassen werden. Stattdessen bedarf es der Entwicklung geeigneter Verfahren und Institutionen, um diese Kriterien kontextspezifisch und demokratisch, gegebenenfalls immer wieder neu auszuhandeln. Je nach Anwendungskontext und Sensibilität des einzusetzenden Systems kann die Beteiligung der Öffentlichkeit erforderlich sein. Dabei sollte der Schutz der jeweils bedürftigsten bzw. von Entscheidungen besonders betroffenen Gruppen besonders berücksichtigt werden.

134) Das *zehnte Querschnittsthema* greift Fragen von Transparenz und Nachvollziehbarkeit sowie von Kontrolle und Verantwortung auf. Die häufige Undurchschaubarkeit von KI-Systemen hat verschiedene Ursachen, die vom Schutz geistigen Eigentums über die Komplexität und Nichtnachvollziehbarkeit der Verfahren bis hin zur mangelnden Durchsichtigkeit von Entscheidungsstrukturen, in die der Einsatz algorithmischer Systeme eingebettet

ist, reichen. Transparenz und Nachvollziehbarkeit algorithmischer Systeme stehen zwar in Zusammenhang mit deren Kontrolle und der Verantwortung für ihren Einsatz, sind für beides aber weder zwingend notwendig noch hinreichend.

» *Empfehlung Querschnittsthema 10:* Es bedarf der Entwicklung ausgewogener aufgaben-, adressaten- und kontextspezifischer Standards für Transparenz, Erklärbarkeit und Nachvollziehbarkeit und ihrer Bedeutung für Kontrolle und Verantwortung sowie für deren Umsetzung durch verbindliche technische und organisatorische Vorgaben. Dabei muss den Anforderungen an Sicherheit und Schutz vor Missbrauch, Datenschutz sowie dem Schutz von intellektuellem Eigentum und Geschäftsgeheimnissen in angemessener Weise Rechnung getragen werden. Je nach Kontext sind hier unterschiedliche Zeitpunkte (ex ante, ex post, Realtime) sowie unterschiedliche Verfahren und Grade der Offenlegung zu spezifizieren.

135) Zusammenfassend geht es in dieser Stellungnahme um die Auswirkungen einer zunehmenden Delegation menschlicher Tätigkeiten an digitale Technologien, insbesondere KI-basierte Softwaresysteme. In zahlreichen Beispielen aus den Bereichen der Medizin, der schulischen Bildung, der öffentlichen Kommunikation und Meinungsbildung sowie der öffentlichen Verwaltung zeigt sich, dass dieses Delegieren sowohl mit Erweiterungen als auch mit Verminderungen menschlicher Handlungsmöglichkeiten einhergeht und sich dadurch sowohl förderlich als auch hinderlich auf die Realisierung menschlicher Autorschaft auswirken kann.

136) Ziel und Richtschnur ethischer Bewertung muss dabei immer die Stärkung menschlicher Autorschaft sein. Dabei ist zu berücksichtigen, dass die Erweiterung von

Handlungsmöglichkeiten für eine Personengruppe mit deren Verminderung für andere einhergehen kann. Diesen unterschiedlichen Effekten ist Rechnung zu tragen, insbesondere im Hinblick auf den Schutz und die Verbesserung der Lebensbedingungen vulnerabler oder benachteiligter Gruppen. Letztlich zeigt sich, dass die normativen Anforderungen an die Gestaltung und den Einsatz solcher Technologien, zum Beispiel in Bezug auf Anforderungen hinsichtlich Transparenz und Nachvollziehbarkeit, den Schutz der Privatsphäre sowie die Verhinderung von Diskriminierung, zwar in allen Bereichen und für alle Betroffenen von hoher Bedeutung sind, sie jedoch sektor-, kontext- und adressatenspezifisch konkretisiert werden müssen, um angemessen zu sein und wirksam werden zu können.

1 EINLEITUNG

Digitale Technologien im Allgemeinen und Systeme sogenannter Künstlicher Intelligenz (KI) im Speziellen durchdringen zunehmend unsere Lebenswelt. Dies reicht von der suchmaschinenbasierten Sortierung von Ergebnissen und Empfehlungen für Filme und Musik über Navigationssoftware bis hin zur Nutzung von Risikoprofilen und Software zur Entscheidungsunterstützung im Sozial- und Justizwesen oder bei der Polizei. Wettervorhersagen und Klimamodellierung, Krebsdiagnostik in der Medizin und psychotherapeutische Erstversorgung mittels Chatbots, intelligente Tutorsysteme zum Vokabellernen oder Emotionserkennung in Videoanalysen – datenbasierte Analysen, Prognosen und das Delegieren von Entscheidungen an Softwaresysteme verdeutlichen, dass digitale Technologien und insbesondere auch KI in nahezu alle Bereiche des öffentlichen und privaten Lebens Einzug gehalten haben. Auch die Debatten um den im November 2022 vorgestellten Chatbot ChatGPT und andere Anwendungen sogenannter *generativer KI*, welche automatisiert neue Inhalte in einer Qualität produzieren, bei der oftmals nicht mehr erkennbar ist, dass diese rein maschinell erstellt wurden, zeigen, dass eine grundlegende Auseinandersetzung mit den Wechselwirkungen zwischen Mensch und Maschine erforderlich ist.

Für die ethische Bewertung solcher Technologien und ihres Einsatzes in verschiedenen Bereichen ist es nötig, nicht nur die Technologien selbst zu verstehen. Es gilt darüber hinaus, auch ihre Wechselwirkungen mit den Personen, die sie verwenden oder von ihrer Anwendung betroffen sind, zu analysieren und den Blick auf gesellschaftliche Effekte zu richten.

Hierbei stellen sich Fragen wie zum Beispiel: Welche – positiven wie auch negativen – Auswirkungen hat das Delegieren von Tätigkeiten, die zuvor Menschen vorbehalten waren, an Maschinen? Werden menschliche Autorschaft und die Bedingungen für verantwortliches Handeln erweitert oder

vermindert? Wie wirkt sich der Einbezug digitaler Technologien und KI auf die Handlungsoptionen verschiedener Beteiligter und Betroffener aus? Wessen Möglichkeiten werden durch den Einsatz erweitert, wessen Möglichkeiten vermindert? Diesen Fragen geht der Deutsche Ethikrat in der vorliegenden Stellungnahme nach und schreibt damit Themen fort, die er bereits in seinen Stellungnahmen „Big Data und Gesundheit – Datensouveränität als informationelle Freiheitsgestaltung" (2017) sowie „Robotik für gute Pflege" (2020) angeschnitten hat. Mit den aktuellen Ausführungen reagiert der Ethikrat zudem auf eine im Oktober 2020 vom Präsidenten des Deutschen Bundestages formulierte Bitte, die Arbeit der Enquete-Kommissionen „Künstliche Intelligenz – Gesellschaftliche Verantwortung und wirtschaftliche, soziale und ökologische Potenziale" und „Berufliche Bildung in der digitalen Arbeitswelt" um eine grundlegende Einbettung des politischen und gesellschaftlichen Diskurses zum Thema KI im Rahmen einer multidisziplinären Stellungnahme zu den ethischen Fragen des Verhältnisses von Mensch und Maschine zu ergänzen.

Grundfrage und Maßstab der hier vorgelegten ethischen Bewertung technologischer Entwicklungen und ihres Einsatzes in verschiedenen Kontexten ist, ob durch die Delegation von Tätigkeiten an Maschinen – bis hin zu einer möglichen Ersetzung – die Bedingungen für verantwortliches Handeln und menschliche Autorschaft erweitert oder vermindert werden.

Die Stellungnahme gliedert sich in drei Teile:

Teil I dieser Stellungnahme liefert die technischen, theoretischen und methodischen Grundlagen. In Kapitel 2 werden in knapper Form die *technologischen Grundlagen* der in dieser Stellungnahme behandelten Entwicklungen dargelegt. Hierzu wird zunächst ein kurzer historischer Überblick über die Entstehungsgeschichte des Forschungsfeldes der KI seit den 1950er-Jahren nachgezeichnet. Gerade in Anbetracht der oft schillernden öffentlichen Debatte rund um sogenannte *starke* oder *generelle* KI werden zentrale Begriffe und Debatten eingeordnet. Im nächsten Schritt werden bedeutsame

Entwicklungen der letzten Jahre skizziert, in denen die Steigerung der Rechenleistung, die Miniaturisierung in der Computertechnik und die zunehmende Vernetzung von Geräten dazu geführt haben, dass unsere Lebenswelt immer stärker von digitalen Technologien durchdrungen ist, was wiederum eine wachsende Datenflut und immer neue Einsatzmöglichkeiten für algorithmische Systeme nach sich zieht. Das Kapitel endet mit einem knappen Verweis auf existierende ethische Leitlinien verschiedenster Akteursgruppen und den aktuellen Rechtsrahmen.

Kapitel 3 widmet sich den *zentralen Begriffen und philosophischen Grundlagen* dieser Stellungnahme. Hierfür werden zunächst zentrale Begriffe beleuchtet, die für das Verständnis sowohl der Unterschiede als auch der Wechselwirkungen zwischen Menschen und Maschinen bedeutsam sind. Dies umfasst zunächst den Begriff der *Intelligenz* selbst und sein Verhältnis zu Begriffen wie *praktischer* und *theoretischer Vernunft*. Von zentraler Bedeutung sind ferner die Begriffe der *Handlung* und der *Verantwortung*. Softwaresysteme werden zunehmend zur Unterstützung von Handlungen und Entscheidungen herangezogen oder diese werden gar vollständig an sie delegiert. In der Folge stellt sich auch die Frage, welche Auswirkungen diese Prozesse für Verantwortungsübernahme und -zuschreibung haben. Das Kapitel endet mit einem Blick auf die anthropologischen Aspekte des Mensch-Maschine-Verhältnisses und beleuchtet zentrale Differenzen zwischen Menschen und Maschinen.

In Kapitel 4 hingegen geht es weniger um die Unterschiede zwischen Menschen und Maschinen als vielmehr um deren vielfältige und mehrstufige Relationen und Wechselwirkungen. Nach einer kurzen Einordnung des im Deutschen Ethikrat verwendeten Verständnisses dieser Wechselwirkungen und dessen Einordnung zwischen den Polen des Technikdeterminismus und des Sozialkonstruktivismus wird mit den Begriffen des *Erweiterns*, *Verminderns* und *Ersetzens* eine Matrix zur Beschreibung und Bewertung von Technikentwicklung und

Technikeinsatz entwickelt. Diese Matrix bildet die Grundlage für die Analysen im folgenden zweiten Teil der Stellungnahme, in dem die Auswirkungen des Delegierens von menschlichen Handlungssegmenten bis hin zu einem vollständigen Ersetzen von Menschen durch Maschinen in verschiedenen Kontexten und Sektoren auf menschliche Autorschaft und die Bedingungen für verantwortliches Handeln untersucht werden.

Teil II dieser Stellungnahme werden die zuvor angestellten Überlegungen anhand von Analysen in vier ausgewählten Anwendungsfeldern exemplarisch konkretisiert: dem Bereich der *Medizin* (Kapitel 5), dem Bereich der schulischen *Bildung* (Kapitel 6), dem Bereich der *öffentlichen Kommunikation und Meinungsbildung* (Kapitel 7) sowie dem Bereich der *öffentlichen Verwaltung* (Kapitel 8). In allen vier Sektoren werden datenbasierte Softwaresysteme eingesetzt. Allerdings unterscheiden sich die Tiefe und Breite der Durchdringung deutlich.

Während die für den Bereich der öffentlichen Kommunikation und Meinungsbildung so zentralen sozialen Medien als Paradebeispiel einer sehr umfassenden Delegation vormals menschlicher Tätigkeiten an Algorithmen dienen können, beispielsweise hinsichtlich der Kuratierung und Moderation von Inhalten, zeigt sich in Deutschland im Bereich der schulischen Bildung noch eine vergleichsweise geringe Nutzung digitaler Technologien im Allgemeinen und KI-basierter Systeme im Speziellen. Eine vollständige Ersetzung menschlicher Tätigkeiten durch Maschinen scheint hier in weiter Ferne. In der Medizin stellt sich dies wiederum anders dar, werden doch zunehmend nicht nur einzelne Bearbeitungsschritte, sondern ganze Funktionen an KI-basierte Softwaresysteme delegiert. Dies reicht von der Nutzung von Softwaresystemen zur Mustererkennung in der Krebsdiagnostik bis zu Chatbots als maschinellen Ersatz von therapeutischem Fachpersonal. Auch in der öffentlichen Verwaltung hat KI in Gestalt datenbasierter Software zur Erstellung von Risikoprofilen und zur Entscheidungsunterstützung Einzug gehalten.

Zentrale Einsicht dieser Analysen ist, dass Entscheidungen über die beste Form und das richtige Ausmaß der Delegation von Tätigkeiten und Funktionen an Softwaresysteme und KI nur kontext-, anwendungs- und personenbezogen spezifiziert werden können. Entsprechend enden alle vier Kapitel mit sektorspezifischen Empfehlungen. Als Richtschnur der Bewertung gilt hierbei jedoch immer,

- ob die Delegation zu einer *Erweiterung* der Handlungsmöglichkeiten, insbesondere zu einer Erhöhung der Möglichkeiten für verantwortliches Handeln und Autorschaft, der verschiedenen Beteiligten und Betroffenen führt oder
- ob es möglicherweise zu einer *Verminderung* von Handlungsmöglichkeiten sowie negativen Auswirkungen auf Möglichkeiten der Autorschaft und Verantwortungsübernahme kommt.

Der Vergleich der Analysen in den vier ausgewählten Anwendungsfeldern erlaubt es, auch übergreifende Themen auszumachen, da es wiederkehrende Aspekte gibt, wenngleich sich diese in den vier Anwendungsfeldern sehr unterschiedlich darstellen. Diese *Querschnittsthemen* und mit ihnen verbundene übergreifende Empfehlungen werden in Teil III der Stellungnahme dargelegt. Dieser letzte Teil beginnt zunächst mit einer Rekapitulation der anthropologischen und ethischen Orientierung, welche dieser Stellungnahme zugrunde liegt, und fasst die Einsichten aus den vorausgehenden Kapiteln im zweiten Teil der Stellungnahme knapp zusammen. Sodann werden die zehn identifizierten Querschnittsthemen dargelegt, die jeweils auch übergreifende Empfehlungen enthalten.

Das erste Querschnittsthema beschäftigt sich mit der für diese Stellungnahme leitenden Grundfrage, welche Auswirkungen das Delegieren von Handlungen an Maschinen auf die Erweiterung oder Verminderung von Handlungsmöglichkeiten von Menschen hat. Hierbei wird deutlich, dass Vor- und Nachteile einer solchen Delegation sich nicht nur zwischen

verschiedenen Sektoren, sondern auch für verschiedene Personengruppen stark unterscheiden. In der Folge muss ein verantwortungsvoller Einsatz von KI diese Nuancen reflektieren und berücksichtigen.

Querschnittsthema 2 adressiert die Auswirkungen von KI auf Wissenserzeugung und den Umgang mit KI-gestützten Voraussagen. Um eine Diffusion von Verantwortung zu verhindern, ist es hierbei von zentraler Bedeutung, KI-gestützte digitale Technologien zur Entscheidungsunterstützung und nicht zur Entscheidungsersetzung einzusetzen.

Querschnittsthema 3 untersucht, inwieweit das Individuum durch statistische Stratifizierung gefährdet ist vor dem Hintergrund, dass es bei datenbasierten Analysen und Prognosen oftmals als Teil eines statistischen Kollektivs behandelt und die Berücksichtigung individueller Aspekte dadurch vernachlässigt wird.

Querschnittsthema 4 beschäftigt sich mit der Frage, welche Auswirkungen das Delegieren von Handlungen an Maschinen auf menschliche Kompetenzen und Fertigkeiten hat und wie der Gefahr von Kompetenzverlusten und Deskilling entgegengewirkt werden kann.

Die Querschnittsthemen 5 und 6 adressieren Chancen und Risiken im Umgang mit Daten. Während es auf der einen Seite darum geht, die Privatsphäre und Autonomie vor übermäßigen Eingriffen und Gefahren durch Überwachung zu schützen, geht es andererseits auch darum, Daten bestmöglich für sinnvolle und gemeinwohlorientierte Nutzung zugänglich zu machen. Hier wird diskutiert, wie Datenschutzrecht und -praxis gestaltet sein müssten, um beiden Zielen optimal Rechnung zu tragen.

In Querschnittsthema 7 wird die Bedeutung digitaler Technologien als kritische Infrastrukturen betont und die Frage gestellt, wie diese sicher und resilient gestaltet und Abhängigkeiten reduziert werden können.

Querschnittsthema 8 knüpft an diese Fragen an und beleuchtet Pfadabhängigkeiten in der Technologieentwicklung

einerseits und Fragen von Missbrauch und Dual Use andererseits.

Die letzten beiden Querschnittsthemen spannen den Bogen noch einmal zurück zu zwei Problemfeldern, die sich in nahezu allen in Kapitel 2 erwähnten ethischen Leitlinien zu KI wiederfinden. Dies betrifft einerseits Fragen rund um systematische Verzerrungen und Diskriminierung (Querschnittsthema 9) sowie andererseits Fragen zur Transparenz, Nachvollziehbarkeit, Kontrolle und Verantwortung im Kontext von KI-Systemen (Querschnittsthema 10). Hier gilt es, kontext-, sektoren- und adressatenspezifische Standards für Transparenz, Nachvollziehbarkeit und die Vermeidung von Verzerrungen zu entwickeln, deren Umsetzung durch verbindliche technische und organisationale Vorgaben gesichert wird.

TEIL I:
TECHNISCHE UND PHILOSOPHISCHE GRUNDLEGUNGEN

2 ZENTRALE ENTWICKLUNGEN UND TECHNISCHE GRUNDLAGEN KÜNSTLICHER INTELLIGENZ

2.1 Historischer Kontext

Die Erfindung von Maschinen und deren Einfluss auf die menschliche Lebenswelt haben sich stets vor dem Hintergrund spezifischer gesellschaftlicher Wahrnehmungen und Erwartungen vollzogen, die Rückwirkungen auf die jeweilige technologische Dynamik und die von ihr ausgehenden Entwicklungen von einfachen Geräten über komplexe Maschinen bis zu großtechnischen Anlagen gehabt haben. Dies gilt auch für die Entstehung immer leistungsfähigerer Softwaresysteme, die in vielfältigen intelligent und autonom anmutenden Formen mit Menschen interagieren können. Auch die dadurch ausgelösten philosophischen und ethischen Fragen und Herausforderungen, mit denen sich der Deutsche Ethikrat in dieser Stellungnahme befasst, sind in diesem historischen Kontext zu betrachten.

Die Idee von Maschinen, deren Fähigkeiten denen des Menschen ähneln oder diese sogar übertreffen, lässt sich Jahrtausende vor Erfindung der ersten Softwaresysteme zurückverfolgen. Schon in den Geschichten der antiken griechischen Mythologie erschufen Götter und Helden animierte Maschinen wie die mechanischen Dienerinnen des Schmiedegottes Hephaistos oder die animierten Statuen des Erfinders Dädalus.[1]

Mit der Erfindung der ersten Computer, also von Maschinen, deren besondere Stärke darin bestand, Programme mit komplexen rechnerischen und logischen Operationen schnell und effizient auszuführen, rückte die Existenz maschineller Intelligenz erstmals in greifbare Nähe. Konzeptionell schon

1 Mayor 2018.

Mitte des 19. Jahrhunderts in den Entwürfen einer „analytischen Maschine“ von Charles Babbage und Ada Lovelace erdacht, wurden die ersten programmierbaren digitalen Rechenmaschinen erst hundert Jahre später gebaut, beginnend mit der von Konrad Zuse und Helmut Schreyer 1941 in Berlin vorgestellten „Zuse Z3“. Parallel dazu entwickelte Alan Turing das formelle mathematische Modell eines universell programmierbaren Computers, später Turingmaschine genannt, das mithilfe von Algorithmen – knapp umreißbar als eindeutig definierte Rechenvorschriften (vgl. Abschnitt 2.2) – aus einer codierten Eingabe[2] eine bestimmte Ausgabe ermittelt.[3]

Turing erkannte früh, dass die Potenziale der neuen Rechenmaschinen auch die Frage aufwarfen, ob und wie es eines Tages möglich sein sollte, dass solche Geräte eine dem Menschen teilweise oder vollständig gleichwertige oder sogar überlegene Intelligenz aufweisen könnten. 1950 formulierte er ein Kriterium für KI, das später als Turing-Test bezeichnet wurde. Demnach sei dann von maschineller Intelligenz auszugehen, wenn das Verhalten einer Maschine für menschliche Beobachter nicht von dem eines Menschen unterscheidbar erscheint.[4] Der Turing-Test gilt als wichtige, wenngleich keineswegs unumstrittene Inspiration für das Forschungsfeld der KI. Der Begriff der Künstlichen Intelligenz selbst wurde in Vorbereitung der Dartmouth-Konferenz geprägt, die im Sommer 1956 über zwei Monate am Dartmouth College im US-Bundesstaat New Hampshire stattfand. Er diente als Sammelbegriff für alle Verhaltensweisen von Maschinen, die man als intelligent bezeichnen würde, wenn Menschen sie zeigen. Dies basiert auf der Prämisse, man könne „jeden Aspekt des Lernens oder jedes andere Merkmal von Intelligenz im Prinzip so genau beschreiben, dass eine Maschine dazu gebracht werden kann, es

2 Zum Beispiel binär als Folge von 0 und 1.
3 Nilsson 2010, 55–59.
4 Turing 1950.

zu simulieren".[5] Im Rahmen der Dartmouth-Konferenz und nachfolgender Konferenzen in den 1950er-Jahren begann man, sich mit vielen Themen zu beschäftigen, die bis heute in der Forschung eine große Rolle spielen, darunter Mustererkennung, Sprachverarbeitung, Abstraktionsfähigkeit, Kreativität, flexibles Problemlösen (z. B. in Strategiespielen wie Schach) und die Fähigkeit zu lernen und sich weiterzuentwickeln.[6]

Die Pionierarbeiten zur KI führten gemeinsam mit rasanten Fortschritten bei der Entwicklung von Computerhardware und Programmiersprachen bald zu großem Optimismus. Viele Forschende gingen davon aus, dass maschinelle Intelligenz in vielen oder sogar allen Facetten innerhalb weniger Jahrzehnte mit menschlicher Intelligenz gleichziehen oder diese sogar übertreffen würde. Zusätzliche Inspiration kam von parallelen Fortschritten in der biologischen Forschung und insbesondere den Neurowissenschaften, deren Ergebnisse zunehmende Einblicke in kognitive Prozesse ermöglichten. Schon zu Beginn der 1970er-Jahre gab es jedoch auch erste Enttäuschungen, weil die in Aussicht gestellten praktischen Erfolge hinter den Erwartungen zurückblieben und Fördermittel daher gekürzt wurden.[7]

Diese erste Skepsis führte dazu, dass der Schwerpunkt der Aufmerksamkeit in den folgenden Jahren stärker auf praktische Anwendungen gerichtet wurde, die zudem dank zwischenzeitlich erfolgten Fortschritten in der theoretischen Grundlagenforschung und Computerentwicklung erstmals in greifbare Nähe rückten. Neben Weiterentwicklungen einzelner Kerngebiete entstanden Bemühungen, diese in anspruchsvolleren Projekten zusammenzuführen. Dazu gehörten etwa erste Versuche, mit parallelen Datenverarbeitungsmethoden

5 Im englischen Original: „[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 et al. 2006, 12).
6 Nilsson 2010, 73–88.
7 Lighthill 1973.

Computer deutlich leistungsfähiger zu machen und sich selbstständig fortbewegende Roboter und Fahrzeuge zu entwickeln. Die zunehmende Vernetzung von Forschenden in Fachgesellschaften und von Computern in Vorläufern des Internets sowie das wachsende Engagement von Forschungsorganisationen, Militär und Industrie trugen dazu bei, dass das Interesse an KI in den 1970er- und 1980er-Jahren erneut aufblühte.[8] Auch hier kam es allerdings zu überzogenen Erwartungen und Voraussagen, denen so viele Enttäuschungen und Mittelkürzungen folgten, dass die Phase ab den späten 1980er-Jahren auch als „KI-Winter" bezeichnet wird.[9] Der anfänglichen Euphorie über viele Entwicklungen folgten Enttäuschungen, wo diese an Grenzen stießen, beispielsweise bei der Leistungs-, Parallelisierungs- und Vernetzungsfähigkeit von Computern sowie der Flexibilität und Lern- bzw. Entwicklungsfähigkeit ihrer Programme.[10]

Parallel zu dieser technischen Entwicklung entstand ein kritischer Diskurs, in dessen Zuge sich Computerethik zunehmend als eigene Disziplin etablierte.[11] Aufbauend auf schon in den 1940ern angestellten Pionierüberlegungen von Norbert Wiener zu gesellschaftlichen Auswirkungen von Interaktionen zwischen Menschen und intelligent agierenden Maschinen[12] hatte in den 1970er-Jahren unter anderen Joseph Weizenbaum[13] detailliertere ethische Analysen zur Einbeziehung von Softwaresystemen und KI in menschliche Entscheidungs- und Lebensprozesse entwickelt. 1985 veröffentlichte Deborah Johnson das erste Lehrbuch zum Thema Computerethik[14] und James Moor betonte die breite Relevanz des Themas. Nach Moor brachte gerade die Flexibilität von Computern als universelle Werkzeuge besondere und weitreichende ethische

8 Nilsson 2010, 343–377.
9 Ebd., 381–429.
10 Schwartz 1986.
11 Bynum 2015.
12 Wiener 1950.
13 Weizenbaum 1976.
14 Johnson 1985.

Herausforderungen mit sich, da sie Menschen ständig neue Handlungsmöglichkeiten erschlossen, für die es oft noch keine geeigneten Regeln oder ethischen Standards gab.[15]

Mit dem Rüstzeug zur systematischen Prüfung von Chancen, Risiken und fundamentalen ethischen Aspekten von KI kamen zunehmend auch philosophische Zweifel auf, ob insbesondere die von einigen Forschenden in Aussicht gestellten Visionen einer generellen oder starken KI jemals realisiert werden könnten – oder sollten.[16] Auch die unkritische Übertragung bzw. Verwendung normativ bedeutsamer Begriffe wie „Intelligenz" im Zusammenhang mit Softwaresystemen geriet in Kritik. Viele Forschende konzentrierten sich fortan bewusst auf enger umrissene Teilbereiche des Felds wie beispielsweise das maschinelle Lernen und versuchten, diese auch sprachlich als eigenständig wahrgenommene Felder zu etablieren, um das mit dem KI-Begriff verbundene Stigma der Aussichtslosigkeit oder moralischen Verwerflichkeit zu vermeiden.

Ab den 1990er-Jahren nahmen drei Entwicklungen Fahrt auf, die der Entwicklung von KI zu neuer Dynamik verhalfen:

Es gab erstens große Fortschritte beim Bau immer leistungsfähigerer, erschwinglicherer und handlicherer Computer, was zu einer massiven Ausweitung ihrer Verbreitung und Nutzbarkeit auf allen Ebenen führte (vgl. Abschnitt 2.2.1). Parallele Datenverarbeitung wurde nach der Jahrtausendwende selbst für viele Privatgeräte Standard und umfasst bei Großrechnern inzwischen bis zu mehrere tausend Prozessoren. Die voranschreitende Miniaturisierung in der Computertechnik und die Entwicklung entsprechend robuster und kleiner leistungsstarker Sensoren erlauben ihren Einsatz in immer mehr Geräten, von denen viele, wie zum Beispiel Smartphones, Tablets, Smartwatches und weitere Wearables, mittlerweile auch mobil nutzbar sind. Viele solcher Geräte können zudem über vielfältige Sensoren Details über Umwelt und die sie nutzenden

15 Moor 1985.
16 Searle 1980.

Personen wahrnehmen, etwa über Kameras, Mikrofone, Bewegungsdetektoren, Ortsdatenempfänger, Thermometer oder Pulsmesser.

Zweitens hat die dynamische Entwicklung des Internets und diverser, auch drahtloser Zugänge die Vernetzungsmöglichkeiten zwischen Geräten revolutioniert, sodass gerade dort, wo viele Alltagsgeräte am Austausch von Daten beteiligt sind, mitunter von einem „Internet der Dinge" die Rede ist.[17] Im Zuge erweiterter Vernetzungsmöglichkeiten sind neue Infrastrukturen zur parallelen und dezentralen Datenverarbeitung entstanden, darunter das sogenannte Cloud-Computing, bei dem Daten aus vielen Quellen zentral verarbeitet und gespeichert werden, oder Edge-Computing, bei dem die Datenverarbeitung zu großen Teilen dezentral auf den Endgeräten erfolgt. Gemeinsam erlaubten diese Entwicklungen immer schnellere, vielfältigere und umfangreichere Verknüpfungen verschiedener, von unterschiedlichen Geräten erfassten, produzierten und (vor-)verarbeiteten Daten, und zwar sowohl zwischen dezentralen Geräten als auch zwischen diesen Geräten und zentralen Hochleistungsrechnern.

Dies hat drittens eine beispiellose und noch immer ansteigende Datenflut hervorgebracht. Der auch unter dem Stichwort „Big Data" zusammengefasste Trend, große Mengen vielfältiger Daten mit hoher Geschwindigkeit[18] zu generieren, zu erfassen, immer wieder neu zu verknüpfen und zu analysieren, hat viele der im Folgenden vorgestellten jüngeren Entwicklungen von Algorithmen vorangetrieben oder überhaupt erst ermöglicht, insbesondere neuere statistische Methoden zur Mustererkennung im Bereich des maschinellen Lernens und hier vor allem das sogenannte Deep Learning.[19]

17 Mattern/Flörkemeier 2010.
18 Im Englischen werden die Kernmerkmale von Big Data auch als „die drei V" – *volume* (Menge), *variety* (Vielfalt) und *velocity* (Geschwindigkeit) – zusammengefasst (Laney 2001).
19 Deutscher Ethikrat 2017, 54 f.

Gemeinsam haben diese drei noch andauernden Trends der Leistungssteigerung und Miniaturisierung von Computern, der Vernetzung digitaler Systeme und der damit verbundenen neuen Möglichkeiten der Datenzusammenführung und -auswertung eine Reihe von nachfolgend näher vorgestellten Entwicklungen angestoßen, deren gesellschaftliche und ethische Analyse im Mittelpunkt dieser Stellungnahme steht.

2.2 KI im 21. Jahrhundert: Big Data, Algorithmen und soziotechnische Ökosysteme

2.2.1 Digitale Durchdringung der menschlichen Lebenswelt

Die im vorigen Abschnitt beschriebene Dynamik der Entwicklungen seit der Jahrtausendwende und insbesondere in den letzten Jahren hat zu einer intensivierten Durchdringung der Alltagswelt mit Computern geführt, die dazu beiträgt, dass unter dem Schlagwort „Künstliche Intelligenz" inzwischen eine Fülle unterschiedlicher, aber vielfach konvergierender Phänomene diskutiert wird. Lag der Fokus angesichts der rasant wachsenden Datenmengen und der damit verbundenen neuen Auswertungsmöglichkeiten zunächst noch auf Big Data, stehen inzwischen vielfach die Leistungen allgegenwärtiger algorithmischer Systeme im Mittelpunkt, die auf breiten Datengrundlagen scheinbar selbst Entscheidungen treffen und so zumindest den Eindruck erwecken, sie würden eigenständig und intelligent agieren.

Vernetzte Computertechnik in Alltagsbegleitern wie Mobiltelefonen, Uhren und Haushaltsgeräten wird nicht nur mit dem Attribut *smart* beworben, sondern wirkt tatsächlich „klug", wenn sie über Sensoren erfasste lokale Gegebenheiten mit online verfügbaren Daten verknüpft, um sich an individuelle Besonderheiten und Bedürfnisse der Menschen, die sie

nutzen, anzupassen. Das gilt erst recht, wenn eine personalisierte oder gar emotional wirkende Ansprache hinzukommt, zum Beispiel über die Stimmen virtueller Assistenzsysteme wie Alexa oder Siri. Den Empfehlungssystemen zahlreicher Internetangebote wie zum Beispiel Videostreamingdienste oder Onlineshops gelingen derweil mittels Analysen von Klicks und Nutzungsprofilen immer treffsichere Vorhersagen über die Vorlieben von Einzelpersonen und auch Roboter und Fahrzeuge können sich dank zunehmend leistungsfähiger Sensorik und Computertechnik immer besser unabhängig von menschlicher Steuerung fortbewegen.

Durch diese vielfältige Verbreitung algorithmischer Technik entstehen soziotechnische Datenökosysteme, in denen über die von Menschen verwendeten Geräte und die sie verknüpfenden Datennetzwerke zunehmend akkurate und umfangreiche digitale Repräsentationen der Bewegungen, Handlungen, Eigenschaften und Präferenzen vieler Personen entstehen. Solche reichhaltigen digitalen Abbilder können nicht nur ausgewertet werden, sondern wirken auch unmittelbar oder mittelbar auf die analogen Prozesse der menschlichen Lebenswelt und menschliches Verhalten zurück, indem auf ihrer Grundlage Menschen Informationen oder Handlungsempfehlungen angeboten werden.

Hinter diesen und vielen weiteren Entwicklungen steht ein Portfolio technischer Entwicklungen, die im Folgenden kurz vorgestellt werden – wohl wissend, dass es sich dabei angesichts der schnellen Veränderungen in der Technikentwicklung und der Einsatzpraxis nur um eine Momentaufnahme handeln kann. Auch wenn in den folgenden Ausführungen Daten, technische Infrastrukturen und Algorithmen konsekutiv beschrieben werden, so ist zu berücksichtigen, dass diese in komplexen Systemen miteinander verbunden sind und ihre Leistungen erst im Zusammenwirken erbringen.

2.2.2 Daten und digitale Infrastrukturen

Das Fundament digitaler Operationen und Interaktionen bilden Daten, die von höchst unterschiedlicher Natur wie Qualität sein können. Es gibt unterschiedliche Datentypen (z. B. binäre, ordinale, metrische oder textliche Daten, Bilder, Musik, Videos) und Datenstrukturen (z. B. Listen, Tabellen), die aus vielfältigen Quellen (z. B. Sensoren, Nutzungsstatistiken, Umfragen) in diversen thematischen Zusammenhängen (z. B. Gesundheitsbereich, Finanzsystem, Verkehr, soziale Netzwerke) stammen können. Daten können auch synthetisch generiert sein, wenn nicht genügend „echte" Daten in der gewünschten Qualität zur Verfügung stehen. Dies ermöglicht es etwa, ein System im Umgang mit Daten zu trainieren, die in vorhandenen Datensätzen aufgrund von Verzerrungen unterrepräsentiert sind, so beispielsweise Daten von Personen bestimmter demografischer oder ethnischer Zugehörigkeit.

Die Qualität eines Datensatzes hängt zum einen davon ab, wie genau, vollständig, aktuell oder detailliert die Daten sind. Zum anderen wird sie von den begleitenden Metadaten beeinflusst, die als Leseanleitung Auskunft über den Kontext und die Semantik (Bedeutung) der Daten geben. Metadaten informieren zum Beispiel über die Herkunft der Daten (Erhebungszeitpunkt und -ort, Art oder genaue Identifikation der Quelle oder des Sensors), die Bedeutung konkreter Zahlencodes oder sonstiger Etiketten im konkreten Datensatz und die Zuordnung zu bestimmten Personen oder Kategorien. Die Kombination und Verknüpfung verschiedener Daten kann nur dann sinnvoll gelingen, wenn Kontext und Semantik für den jeweiligen Zweck hinreichend klar und kompatibel und die Daten auf dieser Grundlage somit interoperabel sind. Besonders gute Interoperabilität kann durch eine klare Standardisierung von Datenformaten und zu erfassenden Metadaten erreicht werden. Dies erhöht die Chancen, dass Daten beispielsweise im Gesundheitssystem institutionenübergreifend analysiert

werden können, selbst wenn sie in verschiedenen Praxen, Kliniken oder Forschungseinrichtungen erhoben wurden.

Die Datenqualität hängt dabei nicht nur von den zuvor genannten Faktoren ab, sondern auch vom Verhältnis zwischen den Erhebungs- und den Anwendungskontexten. Selbst wenn jedes einzelne Datum „korrekt" gemäß der im vorigen Absatz genannten Anforderungen erhoben wurde, haben die Umstände und Zwecke der Datenerhebung wie auch der Datenauswertung Einfluss darauf, wie sehr bestimmte Daten den konkreten Qualitätsanforderungen für einen bestimmten Auswertungszusammenhang genügen (Validität). So werden beispielsweise Daten dafür genutzt, um auf kognitive oder emotionale Zustände von Individuen zu schließen, die nicht direkt beobachtbar sind, wie etwa Aufmerksamkeit, Stress, kognitive Belastung oder Angst. Die Auswahl und Bewertung der Aussagekraft der zu verwendenden Daten sind daher alles andere als trivial und müssen sorgfältig geprüft werden. Es geht also darum, dass Daten nicht nur prinzipiell von hoher oder niedriger Qualität sind, sondern auch für eine jeweilige Frage oder Aufgabenstellung mehr oder weniger angemessen sein können. Werden solche Fragen der Passung nicht rechtzeitig und hinreichend berücksichtigt, sind Verzerrungen oder irreführende Analysen möglich.[20]

Entscheidend für die Leistungsfähigkeit datengetriebener Anwendungen sind auch die Hardware und Infrastruktur, die für die Handhabung und Nutzung von Daten zur Verfügung stehen. Das Herzstück bildet hierbei die Rechenleistung der Prozessorkerne, die mit den Daten arbeiten. Gerade für besonders rechen- und datenintensive Anwendungen kommen inzwischen hochparallelisierte Rechnersysteme zum Einsatz, in denen viele Hunderte oder Tausende Prozessorkerne gleichzeitig arbeiten. Inzwischen gibt es auch Prozessoren, die auf schnelles maschinelles Lernen auf Grundlage großer Datenmengen spezialisiert sind, wie die von Google entwickelten

20 Barocas/Selbst 2016.

Tensor-Prozessoren (Tensor Processing Units). Sie sind für das hochparallele Addieren und Multiplizieren von Matrizen in neuronalen Netzen optimiert und werden beispielsweise in KI-Systemen wie dem Brettspielprogramm AlphaGo oder in den Algorithmen von Google Fotos, Google Maps, der Datenanalyseplattform Kaggle und der Google Cloud Platform verwendet.

Ein Großteil datenintensiver Operationen findet mittlerweile in Großeinrichtungen wie Datenlagern (Data Warehouses, Data Lakes) und Serverfarmen statt, die auf Datenspeicherung und/oder Datenanalyse spezialisiert und durch das Internet untereinander und mit Endgeräten vielfältig und dynamisch vernetzt sind. Der Austausch zwischen den vielen verschiedenen und oft weltweit verteilten Geräten in einem Datennetzwerk funktioniert mithilfe standardisierter Protokolle und Schnittstellen zur Anwendungsprogrammierung (Application Programming Interfaces). Sind Netzwerkknoten erst einmal über solche Schnittstellen verknüpft, fließen Daten entsprechend den vorgenommenen Einstellungen automatisch und oft in Echtzeit. Dies ermöglicht es, Rechen- und Speicherressourcen schnell und flexibel an die Anforderungen bestimmter Projekte und Kunden angepasst zur Verfügung zu stellen. Als Sammelbegriff für solche über das Internet zugänglichen Datenspeicher und Datenanalysedienste hat sich der Begriff des Cloud-Computings etabliert. Der Markt wird im privaten wie im institutionellen Bereich von Angeboten großer Internetfirmen wie Amazon, Microsoft und Google dominiert.

Mit der Leistungssteigerung der Prozessoren in Endgeräten wie Heimcomputern, Mobiltelefonen und Smartwatches wachsen auch die Möglichkeiten, Daten zumindest teilweise bereits lokal in den Geräten, die sie erheben, zu verarbeiten. Solche erhebungsnahen, dezentralen Ansätze zur Datenverarbeitung an den „Rändern“ des Internets werden in Abgrenzung zum Cloud-Computing auch als Edge-Computing bezeichnet. Sie bieten neben Ressourcenschonung beim Datentransfer

auch datenschutzfreundlichere Gestaltungsmöglichkeiten, da zum Beispiel bestimmte sensible Daten gar nicht mehr weitergegeben werden müssen, sondern nur die von ihnen abgeleiteten Ergebnisse der Vorverarbeitung.[21]

2.2.3 Algorithmen und Datenverarbeitung

Mit der Menge und Vielfalt von Daten und der sie miteinander vernetzenden Infrastrukturen wachsen sowohl die Ansprüche an als auch die Möglichkeiten zur Datenverarbeitung. Herzstück jeglicher Datenverarbeitung sind Algorithmen: Verarbeitungsanweisungen zur Lösung eines Problems. Algorithmen geben vor, wie eingegebene Daten meist schrittweise nach klar definierten Regeln umgeformt werden, bis der gesuchte Ausgabewert erreicht ist. In ihrer einfachsten Form können Algorithmen Anleitungen zur klar festgelegten Verarbeitung von Daten oder Informationen sein, zum Beispiel die Formel zur Berechnung des Body-Mass-Indexes[22] oder eine Regel zum Sortieren von Zahlen nach ihrer Größe. Die meisten Algorithmen enthalten konditionale Elemente wie Wenn-dann-Anweisungen oder ineinander verschachtelte Befehlsschleifen, die das Ausführen komplexer Operationen unter Berücksichtigung der jeweiligen Situation ermöglichen. In Computerprogrammen sind Algorithmen in Programmiersprachen codiert.

Im Mittelpunkt der Arbeit mit großen und vielfältigen Datenmengen, die kennzeichnend für moderne soziotechnische Ökosysteme ist, stehen statistische Analysen, mit denen Regelmäßigkeiten in Daten erkannt sowie Zusammenhänge und potenzielle Wirkmechanismen zwischen einzelnen Merkmalen

21 Vgl. etwa die Ansätze des Federated Learnings (Kaissis et al. 2021) und Swarm Learnings (Warnat-Herresthal et al. 2021).

22 Der Body-Mass-Index oder Körpermasseindex wird als Richtwert zur Beurteilung des Verhältnisses zwischen Köpergröße und Körpergewicht verwendet. Er wird berechnet, indem man die Körpergröße (in Metern) durch das Quadrat des Körpergewichts (in Kilogramm) teilt.

identifiziert werden, zum Beispiel Korrelationen zwischen der Körpergröße eines Kindes und der Größe der Eltern, den finanziellen Ressourcen der Familie oder der Ernährung. Ausgehend von solchen Korrelationen können Vorhersagen für ähnliche Datensätze oder künftige Entwicklungen abgeleitet werden. Dabei kommen unterschiedliche statistische Methoden zum Einsatz, die von einfachen Regressionsverfahren bis zu Deep Learning reichen.

Geht es darum, kausale Mechanismen nachzuweisen, sind in der Regel weitere Überlegungen und Untersuchungen nötig, die eine plausible Erklärung für den vermuteten Wirkzusammenhang zwischen Merkmalen anbieten, welche sich auch empirisch – beispielsweise in Experimenten – überprüfen lässt. Eine plausible Hypothese und ein empirischer Nachweis für einen kausalen Zusammenhang zwischen der Anzahl von Störchen und Geburten wird sich wohl nicht etablieren lassen, der Wirkmechanismus eines potenziellen Arzneimittels hingegen schon. Ein solcher Nachweis ist etwa im medizinischen Bereich besonders wichtig, da klinische Studien nur sinnvoll und sicher durchgeführt werden können, wenn biologisch plausible Hypothesen und zumindest vorläufig gesicherte Erkenntnisse zur Wirkweise neuer Therapeutika vorliegen. In anderen Fällen ist die Frage nach Kausalität möglicherweise weniger bedeutsam und es genügt, wenn auf Grundlage vorhandener Daten hinreichend zuverlässige Modelle berechnet werden können. Wenn das Empfehlungssystem eines Onlineshops treffsicher vorhersagt, dass Personen mit bestimmten Merkmalskombinationen unterschiedliche Produkte mögen, reicht das vielleicht für den Einsatzzweck schon aus und die Frage, *warum* solche Zusammenhänge bestehen, muss nicht beantwortet werden.

Statistische Analysen enthalten Unsicherheiten und geben in der Regel an, wie groß die jeweils zu erwartenden Fehler und Unsicherheiten sind, um Ergebnisse angemessen interpretieren zu können. Mit der Menge und Qualität der Daten wächst dabei für gewöhnlich die Verlässlichkeit der Analyseergebnisse,

da die Wahrscheinlichkeit sinkt, dass die beobachteten Muster rein zufallsbedingt sind. Ganz ausmerzen lassen sich Fehler und Unsicherheiten in der Regel allerdings nicht und mit der Minimierung bestimmter Fehlerquellen können andere Fehlerquellen verstärkt werden. Optimiert man beispielsweise einen Test auf eine Virusinfektion dahingehend, dass er durch Berücksichtigung möglichst vieler Merkmale wirklich keine infizierte Person übersieht, steigt das Risiko, auch Personen fälschlich als infiziert zu identifizieren, bei denen vielleicht nur wenige dieser Merkmale in schwacher Ausprägung vorliegen – es entstehen *falsch-positive* Ergebnisse. Optimiert man den Test hingegen so, dass solche Fehler vermieden werden, indem nur wenige und/oder starke Zusammenhänge zwischen Merkmalen berücksichtigt werden, steigt das Risiko, dass tatsächlich infizierte Personen übersehen werden – es entstehen *falsch-negative* Ergebnisse. Welche Fehler in statistischen Analysen am ehesten in Kauf zu nehmen sind, hängt daher auch immer von der konkreten Fragestellung und Zwecksetzung ab und ist in zahlreichen Bereichen nicht nur eine technisch-methodische, sondern eine ethische Frage. Dies zeigt sich beispielsweise bei Entscheidungsunterstützungssoftware im Bereich der Prognose von Kindeswohlgefährdung. Schon im Rahmen der Softwareentwicklung muss entschieden werden, welcher Fehler schwerer wiegt: eine Kindeswohlgefährdung übersehen oder ein Kind unbegründeterweise aus einer Familie genommen zu haben.

Maschinelles Lernen

Für algorithmische Verfahren und Systeme, die ihre Mustersuche, Modellbildung und sonstige Funktionsweise datenbasiert optimieren können, hat sich der Begriff des maschinellen Lernens etabliert.[23] Er umfasst unterschiedliche Ansätze, die in verschiedenen Anwendungsfeldern zum Einsatz kommen

23 Zu Details zu den in diesem Abschnitt kurz vorgestellten Ansätzen maschinellen Lernens vgl. Bauckhage et al., in: Görz/Schmid/Braun 2021, 429 ff.

und ständig weiterentwickelt werden. Gemeinsam ist diesen Ansätzen eine anfängliche Trainingsphase, in der ein Algorithmus sein Modell zur Mustererkennung durch wiederholte Analyse von Trainingsdaten aufbaut und verfeinert. In der öffentlichen Debatte wird der Begriff häufig synonym mit KI verwendet. Allerdings beschreibt er lediglich eine Reihe statistischer Verfahren zur Analyse großer Datenmengen, die von gut interpretierbaren Ansätzen wie Entscheidungsbaumalgorithmen über statistische Optimierungsmethoden wie Support Vector Machines bis hin zu künstlichen neuronalen Netzen reichen.[24] KI hingegen beschreibt – wie in Kapitel 2 bereits dargelegt – ein breites Forschungsfeld, in dem bereits seit den 1950er-Jahren ganz unterschiedliche Methoden eingesetzt werden, um menschliche Kognition maschinell zu simulieren. Der aktuelle „KI-Sommer" geht wesentlich auf rasante Weiterentwicklungen beim maschinellen Lernen zurück. Waren klassische Ansätze noch stark auf passend vorverarbeitete Daten angewiesen, können moderne Methoden insbesondere des Deep Learnings (siehe unten) flexibel auf verschiedenste Daten wie Bilder, Videos und Texte angewendet werden.[25]

Im Kontext des maschinellen Lernens sind folgende Typen von Lernverfahren zu unterscheiden: überwachtes Lernen, unüberwachtes Lernen und Verstärkungslernen. Beim *überwachten Lernen* sind die Zuordnungen zwischen den Eingabe- und den gesuchten Ausgabedaten im Trainingsdatensatz bereits bekannt. In einem Trainingsdatensatz zur Erkennung von Hautkrebs wären dies zum Beispiel Bilder von gesunder Haut und Hautkrebs, deren gesicherte Zuordnung zu einer dieser beiden Kategorien in einem Etikett (Label) vermerkt ist. Der trainierende Algorithmus kann seinen Zuordnungserfolg anhand dieser Etiketten in jeder Trainingsrunde überprüfen und seine Rechenregeln anpassen, bis das System für die korrekte

24 Schmid 2022.

25 Dazu gehören beispielsweise Convolutional Neural Networks, Long Short-Term Memory und generative Ansätze (Bauckhage et al., in: Görz/Schmid/Braun 2021, 509 ff.).

Erkennung von Hautkrebs optimiert ist. Überwachtes Lernen eignet sich besonders für Algorithmen, die bestimmte Muster in Daten zuverlässig erkennen und kategorisieren können sollen, also zum Beispiel für Diagnostikwerkzeuge, die wie im obigen Beispiel in der Bilderkennung eingesetzt werden, aber auch für Ansätze zur Sprach-, Text- oder Objekterkennung. Auch das Training von Algorithmen zur Vorhersage künftiger Entwicklungen, beispielsweise beim Wetter, im Finanzsektor oder beim Wasser- oder Stromverbrauch, geschieht häufig mit überwachtem Lernen, da Algorithmen hier Regressionsmodelle auf Grundlage bekannter und entsprechend etikettierter Zusammenhänge zwischen Merkmalen in Trainingsdaten aus der Vergangenheit erlernen.

Unüberwachtes Lernen hingegen funktioniert ohne vorherige Etikettierung der Trainingsdaten; stattdessen „sucht" der Algorithmus eigenständig nach Mustern. Auf diese Weise können neue Strukturen und Gruppierungen (Cluster) in Daten entdeckt und weiterverwendet werden, zum Beispiel zur Generierung neuer „kreativer" Inhalte durch den Algorithmus oder zur Analyse von Marktsegmenten oder Zielgruppen, um passgenaue Angebote zu liefern. Durch den Verzicht auf eine vorherige Etikettierung lassen sich beim unüberwachten Lernen zudem nicht nur die im Vorfeld notwendigen Vorverarbeitungsschritte, sondern auch bestimmte Verzerrungen (Bias) reduzieren, so zum Beispiel jene, die bei einer händischen Zuweisung von Merkmalen aufgrund von Vorurteilen oder bloßen Intuitionen seitens des Entwicklungsteams einfließen können. Andererseits sind unüberwachte Lernalgorithmen anfällig für jene Verzerrungen, die bereits in den Trainingsdaten enthalten sind. Ein Algorithmus, der Vorhersagen zur Medikamentenverträglichkeit mit Daten trainiert, die überwiegend von Männern stammen, ist so möglicherweise weniger präzise, wenn es um Vorhersagen für Frauen geht, da relevante Faktoren für diese Gruppe beim Training nicht berücksichtigt wurden. Außerdem steigt in dem Maße, in dem ein Algorithmus unabhängig von menschengemachten

Etiketten und sonstigen Vorgaben seinen eigenen Weg zur Problemlösung „sucht“[26], auch die potenzielle Undurchdringbarkeit seiner Lösungsansätze.

Beim *Verstärkungslernen* optimiert der Algorithmus seine Operationen auf bestimmte Ziele hin und erhält dabei in der Trainingsphase für jeden Versuch eine Rückmeldung, ob dieser Schritt das System dem Ziel nähergebracht oder es davon entfernt hat. Schritte in die richtige Richtung werden mit einer Bonusfunktion belohnt, solche in die falsche Richtung bestraft. Die Methoden sind in der Regel so aufgebaut, dass sich kurzfristig auch Schritte in die falsche Richtung lohnen können, wenn sie einer langfristig günstigen Strategie dienen. Verstärkungslernen lässt sich beispielsweise einsetzen, um Spielstrategien zu entwickeln, Verkehrsflüsse zu optimieren oder Klicks und Verweildauer auf einer Plattform zu maximieren.

Die algorithmischen Strategien, die im Laufe des Trainings zur Bewältigung der jeweiligen Aufgaben entwickelt werden, sind in der Regel selbst für geschultes Personal, das den Code einsehen kann, nicht vollständig nachvollziehbar (Blackbox). Das gilt umso mehr, je dynamischer sich ein Algorithmus im Laufe seiner Arbeit weiterentwickelt, also insbesondere für Deep Learning. Es gibt verschiedene Lösungsansätze, um trotzdem eine für die jeweilige Zielgruppe angemessene Transparenz, Interpretierbarkeit oder Erklärbarkeit algorithmischer Prozesse zu erreichen (Explainable AI); deren Auswahl und Anwendung ist jedoch technisch anspruchsvoll und stellt ein hochdynamisches Forschungsfeld dar.[27]

26 Wir verwenden gelegentlich mentale Prädikate, um Eigenschaften von Software zu beschreiben. Das entspricht einer etablierten Redeweise, die nicht dahingehend missverstanden werden darf, dass damit angenommen wird, Softwaresysteme hätten tatsächlich diese mentalen Eigenschaften, würden zum Beispiel etwas suchen.

27 Samek et al. 2021; Miller 2019; siehe auch https://facctconference.org [27.02.2023].

Der Begriff des maschinellen Lernens[28] kommt nicht von ungefähr, dienen menschliche Lernprozesse doch als Analogie – ähnlich wie beim Begriff der Künstlichen Intelligenz auch. So lernen auch Menschen „überwacht", wenn sie beispielsweise unterrichtet werden oder sich von anderen etwas abschauen. Unüberwachtes maschinelles Lernen weist Parallelen zu Spiel und kreativen oder erkundenden Prozessen auf. Verstärkungslernen findet biologische Vorbilder in den gut erforschten Belohnungs- und Aversionssystemen von Gehirnen[29], die unmittelbar auf Lernprozesse zurückwirken, indem sie über die Ausschüttung diverser Neurotransmitter Verbindungen zwischen Nervenzellen stärken oder schwächen.

Biologisch inspiriert ist auch ein Teilbereich des maschinellen Lernens, der besonders für den Umgang mit großen Datenmengen geeignet ist und in den letzten Jahren zu einem wichtigen Treiber für viele KI-Anwendungen geworden ist – Deep Learning. Hier kommen sogenannte neuronale Netze zum Einsatz, deren Funktionsweise entfernt an Netzwerkstrukturen im Gehirn angelehnt ist (siehe Infokasten 1).

Infokasten 1: Neuronale Netze im Gehirn

Im Gehirn ermöglicht das Zusammenspiel von Nervenzellen in komplexen und hierarchischen Vernetzungen vielfältige und anpassungsfähige Funktionen. Eine einzelne Nervenzelle kann über Synapsen mit mehreren tausend anderen Nervenzellen verschaltet sein, und zwar sowohl in ihrem (postsynaptischen) Eingabebereich, in dem sie Signale empfängt, als auch in ihrem (präsynaptischen) Ausgabebereich, über den sie Informationen an „flussabwärts" liegende Zellen weitergibt. Bei der Verarbeitung der vielfältigen empfangenen Signale leistet schon die einzelne Zelle erstaunliche Integrationsarbeit, die zum Beispiel die Qualität und die raumzeitlichen Muster der eingehenden Informationen berücksichtigt. Wesentlicher Bestandteil von Lernprozessen sind plastische Veränderungen in der Funktion und Struktur von Synapsen, die in Reaktion auf neuronale Aktivität erfolgen. Eine große

28 Geprägt wurde der Begriff „Machine Learning" von KI-Forschern, die die ersten Ansätze hierzu etabliert haben (Samuel 1959; Carbonell/Michalski/Mitchell 1983). Parallel hat sich im Bereich der Signalverarbeitung die Mustererkennung (Pattern Recognition) als verwandtes Forschungsgebiet etabliert. Beides wird heute als maschinelles Lernen zusammengefasst.

29 Neftci/Averbeck 2019.

Rolle spielt dabei die Modulation des Verhältnisses von synaptischer Erregung und Hemmung innerhalb der neuronalen Netzwerke. Die Stärke der Synapsen wird durch die Anzahl verschiedener Rezeptoren bestimmt, die auf unterschiedliche chemische Botenstoffe (Neurotransmitter) reagieren. Lernergebnisse und neue Gedächtnisinhalte schlagen sich in der Stärkung einzelner Synapsen nieder, was wiederum Rückwirkungen auf die Funktion der beteiligten Netzwerke und sogar auf das Überleben einzelner Nervenzellen haben kann.

Die Bausteine eines maschinellen neuronalen Netzes sind künstliche „Nervenzellen" – Recheneinheiten, die auch als Perzeptrone bezeichnet werden. Jedes Perzeptron ist mit anpassbaren Gewichtungen und einem Schwellenwert ausgestattet, die mitbestimmen, wie das Perzeptron auf eingehende Signale reagiert und welchen Ausgabewert es produziert.

Die Perzeptrone sind miteinander vernetzt, wobei die Vernetzung meist als vollständige Verbindung aller Neuronen einer Schicht mit allen der nächsten Schicht realisiert wird. Jedes neuronale Netz besteht aus einer Eingabeschicht, einer Ausgabeschicht und dazwischenliegenden „versteckten" Schichten, in denen die Datenverarbeitung stattfindet. Es wurden verschiedene Arten von Architekturen neuronaler Netze entwickelt, die für die Bearbeitung bestimmter Arten von Daten besonders gut geeignet sind. Eine der einflussreichsten Architekturen im Bereich des Deep Learnings sind Convolutional Neural Networks. Während klassische neuronale Netze wie auch die meisten anderen Ansätze des maschinellen Lernens Merkmale als Eingabe erwarten, können Convolutional Neural Networks Rohdaten wie Bilder direkt verarbeiten. Man spricht hier von Ende-zu-Ende-Lernen. Dies wird ermöglicht, indem spezielle, mit Filtern ausgestattete Schichten eingeführt werden, mittels derer eine bestimmte Information aus den Bildern herausgefiltert werden kann. Mathematisch entspricht dies der Operation der Faltung (*convolution*). Dabei werden nach und nach bessere Datenrepräsentationen aufgebaut. In der Bilderkennung etwa entwickelt das Netzwerk zunächst Kontrastfilter, die helle und dunkle Bereiche voneinander

unterscheiden, dann Filter, die beispielsweise Konturen oder Ecken erkennen, und schließlich Filter für komplexere visuelle Komponenten (ein Gebäude oder ein Rad) und die kompositionelle Repräsentation ganzer Gegenstände oder Szenen. Für verschiedene Arten von Daten (Text, Bild, Messreihen) und Problemen (Klassifikation, Segmentierung, Generierung) gibt es verschiedene Netzwerkarchitekturen, die sich insbesondere in den letzten Jahren entwickelt oder weiterentwickelt haben.

Im Unterschied zum menschlichen Gehirn sind künstliche neuronale Netzwerke jedoch bislang trotz ihrer enormen Komplexität und ihrer beachtlichen Datenverarbeitungskapazitäten auf vergleichsweise enge Aufgabenbereiche ausgerichtet und auch nach wie vor deutlich schlichter. Der 2018 in Betrieb genommene SpiNNaker-Supercomputer hat über eine Million Rechenkerne, von denen jeder bis zu 256 Nervenzellen nachstellen soll.[30] Im menschlichen Gehirn sind hingegen ca. 86 Milliarden Nervenzellen vernetzt.

Leistungsstarke, auf die Anforderungen neuronaler Netze spezialisierte Hardware zusammen mit neuen Methoden und Architekturen erlaubt es, besser mit großen Datenmengen und mit unstrukturierten Daten umzugehen, da sie die damit einhergehenden Komplexitäten aufgrund ihrer modularen und hierarchischen Struktur besonders gut handhaben und abbilden kann. Der aktuelle Boom bei vielen KI-Entwicklungen geht entscheidend auf die durch Deep-Learning-Algorithmen eröffneten Möglichkeiten zurück und diese wiederum auf die enorm gestiegenen Datenmengen und Chipleistungen.

2.2.4 Einsatzbereiche algorithmischer Systeme und KI

Die oben beschriebenen kombinierten Entwicklungen von Hardware und Software, Vernetzung und Datenproduktion

30 Garside/Plana 2020, 46.

haben vielfältig einsetzbare Anwendungsmöglichkeiten von algorithmischen Systemen hervorgebracht, die auch immer wieder auf großes öffentliches Interesse treffen.

So gelingt es Computern inzwischen, Menschen in anspruchsvollen Strategiespielen wie Schach und Go zu schlagen. Während der Schachcomputer Deep Blue den damaligen Schachweltmeister Garri Kasparow 1997 noch mithilfe eines Algorithmus besiegte, der im Wesentlichen mittels schneller Suche in einer riesigen Datenbank möglicher Spielzüge triumphierte, meistert der 2019 vorgestellte Algorithmus MuZero inzwischen Spiele wie Go, Schach und Shogi ganz ohne die Berücksichtigung historischer Partien oder Kenntnis der Spielregeln. Dieser Algorithmus lernt vielmehr durch Versuch und Irrtum, indem er gegen sich selbst spielt und ausprobiert, was dabei am besten funktioniert.[31]

Im Bereich der Sprachverarbeitung erregen derweil regelmäßig Leistungssprünge in der Textproduktion Aufsehen. Das neuronale Netz des Generative Pre-trained Transformer 3 (GPT-3) des US-amerikanischen Unternehmens OpenAI etwa wurde auf Basis einer ausnehmend umfassenden Menge an Texten – darunter die komplette englischsprachige Wikipedia – trainiert, bis es ab 2020 selbst Texte produzieren konnte, deren maschineller Ursprung oftmals nicht mehr zu erkennen war.[32] Das auf GPT-3 aufbauende, im November 2022 veröffentlichte Dialogsystem ChatGPT[33] kann auf unterschiedlichste Fragen mitunter so überzeugend und differenziert reagieren, dass sich selbst Antworten auf komplexe Aufgaben, wie die Erstellung wissenschaftlicher Hausarbeiten, nicht von qualitativ hochwertigen menschlich verfassten Eingaben unterscheiden lassen.[34]

Der Deutsche Ethikrat nimmt in dieser Stellungnahme vier Handlungsfelder in den Blick, in denen der Einsatz

31 Schrittwieser et al. 2020.
32 Brown et al. 2020.
33 https://openai.com/blog/chatgpt [10.02.2023].
34 Pretschner et al. 2023.

algorithmischer Systeme und Künstlicher Intelligenz entweder schon besonders weitreichende Veränderungen mit sich gebracht hat oder dies in näherer Zukunft bewirken könnte. Im Gesundheitsbereich (vgl. Kapitel 5) beispielsweise stellt die Auswertung großer Datenmengen durch maschinelles Lernen Fortschritte bei Diagnostik und individualisierten Präventions- und Therapieempfehlungen in Aussicht[35] und kommt KI-gestützte Software auch in Robotern zunehmend zur Anwendung, etwa in der Pflege[36], bei Operationen oder in der Psychotherapie. Im Bereich der Bildung (vgl. Kapitel 6) gibt es nicht erst seit der Coronapandemie vielfältige Ansätze, die Vermittlung von Wissen und Kompetenzen in der Schule mithilfe datenbasierter, KI-gestützter Lehr- und Lernsysteme und KI-gestütztem Unterrichtsmanagement effektiver zu gestalten und besser auf individuelle Belange von Lernenden einzugehen. Besonders weit in den Alltag vieler Menschen sind derweil Entwicklungen im Bereich der öffentlichen Kommunikation und Meinungsbildung (vgl. Kapitel 7) vorgedrungen, wo ein Großteil des Informationsaustauschs inzwischen über algorithmisch gestützte digitale Plattformen und soziale Medien abläuft. Dabei können durch Verknüpfung vielfältiger Daten zunehmend präzise Profile von Einzelpersonen erstellt und zur Entwicklung maßgeschneiderter kommerziell, aber auch politisch motivierter Ansprachen und Angebote verwendet werden. Auch in der öffentlichen Verwaltung (vgl. Kapitel 8) kann der Einsatz algorithmischer Systeme das Leben vieler Menschen berühren, beispielsweise bei der Beurteilung oder Überwachung von Personen im Bereich des Sozial- oder Polizeiwesens.

35 Deutscher Ethikrat 2017.
36 Deutscher Ethikrat 2020a.

2.2.5 Ethische Leitlinien und regulativer Rahmen für algorithmische Systeme und KI

Die vorstehend beschriebenen Entwicklungen von Datenanalysemethoden und Softwaresystemen sowie die damit verbundenen Veränderungen in vielen Bereichen menschlichen Handelns bringen auch Herausforderungen für die Ausgestaltung von Regeln für das menschliche Miteinander mit sich. In diesem Zusammenhang ist bereits eine Reihe von Regularien entstanden oder aktuell in der Entwicklung, die hier nur kurz benannt werden.

Ethische Leitlinien

Die potenziell weitreichenden Rückwirkungen datengestützter algorithmischer Anwendungen auf viele Bereiche der menschlichen Lebenswelt haben in der Industrie wie auch im akademischen und zivilgesellschaftlichen Diskurs Überlegungen ausgelöst, wie ein ethisch vertretbarer und gesellschaftlich verträglicher Einsatz solcher neuen Technologien gestaltet werden kann. Dieser Austausch hat eine Fülle an Leitlinien hervorgebracht – 84 allein nach einer Übersicht aus dem Jahr 2019.[37] Das Angebot reicht von Kodizes einzelner Unternehmen wie beispielsweise der Deutschen Telekom[38], von SAP[39], Microsoft[40] oder Google[41] über Richtlinien von Fachgesellschaften wie dem Institute of Electrical and Electronics Engineers[42] bis hin zu Werken auf nationaler oder internationaler Ebene. In Deutschland sind hier insbesondere die Stellungnahmen der Datenethikkommission der Bundesregierung[43] und

37 Jobin/Ienca/Vayena 2019.
38 https://www.telekom.com/en/company/digital-responsibility/details/artificial-intelligence-ai-guideline-524366 [18.01.2023].
39 https://www.sap.com/products/artificial-intelligence/ai-ethics.html [18.01.2023].
40 https://www.microsoft.com/en-us/ai/our-approach [18.01.2023]; https://www.microsoft.com/cms/api/am/binary/RE4pKH5 [01.02.2023].
41 https://ai.google/principles [18.01.2023].
42 IEEE 2019.
43 Datenethikkommission der Bundesregierung 2019.

der Enquete-Kommission Künstliche Intelligenz des Deutschen Bundestages[44] zu nennen, auf internationaler Ebene die „Ethik-Leitlinien für eine vertrauenswürdige KI" der von der Europäischen Kommission eingesetzten Hochrangigen Expertengruppe für künstliche Intelligenz[45] und die „Recommendation on the Ethics of Artificial Intelligence" der UNESCO[46]. Es gibt inzwischen mehrere Ansätze, solche Richtlinien zu kartieren und aus ihren überlappenden Inhalten sowohl Gemeinsamkeiten als auch Unterschiede herauszuarbeiten, die sich aus spezifischen Perspektiven und Interessen der beteiligten Personen und Institutionen ergeben.[47]

Dabei wurden eine Reihe von Themen identifiziert, die weitgehend anwendungsbereichsübergreifend bedeutsam erscheinen und daher in den meisten ethischen Leitlinien zu algorithmischen Systemen und KI eine Rolle spielen. Zu diesen gehören Bedenken hinsichtlich des Schutzes der Privatsphäre ebenso wie Fragen, die sich im Zusammenhang mit der Entwicklung und Funktion von Algorithmen ergeben, beispielsweise zu Voreingenommenheiten und Verzerrungen oder zu Transparenz und Verantwortung angesichts weitreichender Undurchschaubarkeiten von Ansätzen maschinellen Lernens. Versuche, die in diesen Kontexten aufgestellten normativen Forderungen übergeordneten ethischen Prinzipien zuzuordnen, ergeben teilweise eine enge Anlehnung an die etablierten vier Prinzipien der Medizinethik[48] – Autonomie, Schadensvermeidung, Wohltätigkeit, Gerechtigkeit –, ergänzen diese aber häufig um weitere Prinzipien wie Erklärbarkeit und menschliche Kontrolle, die sich auf spezifischere Aspekte von Digitalisierung und KI beziehen.

44 Deutscher Bundestag 2020.
45 Hochrangige Expertengruppe für künstliche Intelligenz 2019.
46 UNESCO 2021.
47 Fjeld et al. 2020; Floridi et al. 2018; Hagendorff 2020; Jobin/Ienca/Vayena 2019; Rudschies/Schneider/Simon 2021.
48 Beauchamp/Childress 2001.

Regulativer Rahmen

Für das Verständnis des regulativen Rahmens ist hilfreich zu sehen, dass Regeln neue Entwicklungen – auch Bereich des Zusammenwirkens von Mensch und Maschine – auf unterschiedliche Weise erfassen können. Abstrakte, generelle Regeln können auf ein neues Phänomen angewandt werden, ohne dass es nötig ist, sie in ihrem Wortlaut zu verändern oder neue Regeln für das Phänomen zu schaffen. Bei gesetzlichen Regelungen ist es Aufgabe von Rechtswissenschaft und Rechtspraxis, die Regeln entsprechend zu interpretieren.

Vor diesem Hintergrund sind für die hier beschriebenen soziotechnischen Entwicklungen beispielsweise Rechtsnormen relevant, die gar keinen expliziten Technikbezug haben. So kann sich etwa aus den allgemeinen Regeln des Kartellrechts ergeben, ob es eine verbotene Preisabsprache darstellt, wenn von zwei Unternehmen eingesetzte Preisfindungsalgorithmen ohne menschliche Steuerung so miteinander interagieren, dass die Preise nicht mehr unabhängig festgelegt werden.[49] Aus generellen Vorschriften des Arbeitsrechts können sich Grenzen dafür ergeben, inwieweit der Arbeitgeber die Leistung von Mitarbeitenden mithilfe technischer Systeme erfassen darf.[50] Technikneutrale Arbeitsschutzbestimmungen sind auch auf den Einsatz von Robotern in der Produktion anwendbar.[51] Viele weitere Beispiele ließen sich hier anführen.[52]

Zu diesen technikunspezifischen, aber hoch relevanten Normen gehören Haftungsregelungen. Sie prägen die Entwicklungen im Bereich Mensch/Maschine entscheidend.[53] Haftungsregeln legen fest, wer im Falle eines Schadens dafür (finanziell) einzustehen hat, unter welchen Voraussetzungen und in welchem Umfang. Sie finden sich als bereichsspezifische

49 Truby/Brown 2021.
50 Holthausen 2021; Waas 2022.
51 Kollmer, in: Kollmer/Klindt/Schucht 2021, Überblick Vor § 1 Rn. 90 ff.; Günther/Böglmüller, in: Arnold/Günther 2022, § 4 Rn. 125–129.
52 Im Überblick: Schulz/Schmees 2022.
53 Yeung 2018.

und bereichsübergreifende Haftungsregelungen. Auch wenn sie nur regeln, wer im Einzelfall nachträglich für Schäden aufkommen muss, wirken sie auf die Herstellung und auf die Anwendung technischer Systeme zurück. Wenn eine Person weiß, dass sie für die Schäden haftet, die das System anrichtet, so hat sie beispielsweise einen Anreiz, das System intensiver zu überwachen, seinen Aktionsradius einzuschränken oder eine menschliche Letztentscheidung vorzusehen. Knüpft die Haftung aber gerade an die menschliche Entscheidung an, kann der Anreiz andersherum auch gerade dahingehen, die Maschine selbstständig „laufen zu lassen". Es kommt also auf die Ausgestaltung der Haftungsregelungen an.

Eine – bislang theoretische – Diskussion zur Haftung betrifft die Frage, ob nicht eine Situation eintreten kann, in der Maschinen so unabhängig von Menschen agieren, dass es unangemessen erscheint, den Menschen, der die Maschine entwickelt oder sie eingesetzt hat, für die Schäden, die die Maschine anrichtet, haften zu lassen. Wenn die Maschine selbst nicht auf Schadenersatz in Anspruch genommen werden kann, entsteht eine Verantwortungs- und Haftungslücke (Responsibility Gap) – Geschädigte gingen also leer aus. Eine rechtliche Antwort darauf kann sein, den Maschinen selbst – oder durch eine Versicherung – Vermögen zuzuordnen, auf das dann im Schadensfall zurückgegriffen werden kann.[54] Dies verweist auf eine noch grundsätzliche rechtliche Diskussion, nämlich ob es irgendwann sinnvoll erscheint, technischen Systemen eine eigene Rechtspersönlichkeit zuzuweisen, wie man es seit Langem bei Organisationen tut (sogenannte juristische Personen).[55]

Neben oder anstelle der Anwendung abstrakter, genereller Regeln können sich im politischen Prozess für neue Phänomene spezielle Normen herausbilden. Als historisches Beispiel kann das Straßenverkehrsrecht dienen, das auf die

54 Denga 2018; Burchardi 2022.
55 Riehm 2020.

Herausforderungen reagierte, die durch Mobilität mit Kraftfahrzeugen im Vergleich zu Pferdekutschen entstanden. Derartige Regeln unterscheiden sich im Hinblick darauf, wie explizit sie an der Technik selbst ansetzen und wie bereichsspezifisch sie sind. Die Regelungskonzepte sind vielfältig.

Nur verhältnismäßig wenige deutsche oder europäische Regelungen nehmen direkt auf Phänomene, wie sie in dieser Stellungnahme beschrieben werden, Bezug. Zu den Ausnahmen zählt die Spezialmaterie des deutschen Finanzmarktrechts, in der das selbstständige Bieten durch algorithmische Systeme geregelt und begrenzt wird, auch um zu verhindern, dass es zu einem von Computern ausgelösten Finanzcrash kommt.[56] Ohne das Wort „Algorithmen" zu erwähnen, wurden die Mediengesetze in Deutschland auf Vielfaltsrisiken bei der Auffindbarkeit von Medieninhalten erweitert, die durch Selektieren und Aggregieren durch sogenannte Intermediäre ausgehen – gemeint sind Plattformen wie Facebook oder Suchmaschinen wie Google (§ 2 Abs. 2 Nr. 16, §§ 91–96 MStV). Diese Intermediäre müssen unter anderem ihre zentralen Kriterien der Sammlung, Auswahl und Darstellung von Inhalten und deren Gewichtung offenlegen (Transparenzgebot); zusätzlich gilt ein Diskriminierungsverbot. Auch in anderen Bereichen wurde und wird auf technische Entwicklungen und Veränderungen durch Gesetzgebung reagiert.[57]

Da die Gefahren oft von Akteuren wie international agierenden IT-Unternehmen ausgehen, gewinnt die Regelung auf transnationaler und insbesondere europäischer Ebene zunehmend an Bedeutung. Relevanz hat vor allem die Datenschutz-Grundverordnung (DSGVO) erlangt, die seit 2018 einen einheitlichen europäischen Rahmen für den Schutz personenbezogener Daten zur Verfügung stellt. Art. 22 DSGVO enthält das Recht, nicht einer Entscheidung unterworfen zu sein, die ausschließlich auf einer automatisierten Datenverarbeitung

56 Kurth, in: Ebers et al. 2020, § 14.
57 Im Überblick: Guckelberger 2019; Hoffmann-Riem 2022.

beruht; es gibt allerdings Ausnahmen.[58] Diese Regelungen sind zu einem Referenzpunkt der Diskussion über automatisiertes Entscheiden geworden, da auch grundsätzliche Fragen wie das Recht auf Eingreifen durch einen Menschen (Human-in-the-Loop)[59] und auf aussagekräftige Informationen zur involvierten Logik[60] vorgesehen sind.

Spezifische Gefahren Künstlicher Intelligenz will die Europäische Union mit einem neuen Rechtsakt regeln, dem Artificial Intelligence Act.[61] Der Gesetzgebungsprozess zeigt allerdings auch, wie schwer es ist, Phänomene wie KI für die Rechtsanwendung handhabbar zu definieren und adäquate Schutzkonzepte (etwa abgestuft nach Risiken) zu entwickeln. Rechtlich nicht bindende Empfehlungen gibt es in diesem Bereich bereits.[62] Ein weiteres Gesetzgebungspaket hat Daten als Referenzpunkt: Der Data Governance Act trat im Juni 2022 in Kraft und schafft Prozesse, Strukturen und einen Rechtsrahmen für die gemeinsame Nutzung von personenbezogenen und nicht personenbezogenen Daten. Ein weiter gehender Data Act ist in Vorbereitung. Auch das Gesetzespaket für digitale Dienste, das die Europäische Union 2022 erlassen hat, bestehend aus dem Digital Markets Act und dem Digital Services Act, hat im vorliegenden Zusammenhang Bedeutung. Beide Gesetze stellen Regeln für die großen Plattformen und Gatekeeper der digitalen Ökonomie auf, die auch im Bereich der Entwicklung und Anwendung von technischen Systemen führend sind.[63]

Der Regelungsrahmen wird zentral durch rechtliche Normen wie die genannten (und zahlreiche weitere) bestimmt, aber nicht vollständig. Selbstregulierung vor allem durch die Industrie (etwa durch oben bereits erwähnte freiwillig

58 Laue 2016.
59 Buckley 2021.
60 Kumkar/Roth-Isigkeit 2020.
61 Bomhard/Merkle 2021; Geminn 2021.
62 Kettemann 2022; Ad Hoc Committee on Artificial Intelligence 2021; OECD 2019.
63 Eifert et al. 2021; Gielen/Uphues 2021.

erlassene Kodizes) ist ebenso Teil des Regelungsrahmens, wie es beispielsweise Verträge etwa zwischen Plattformen und Nutzenden sind, in denen wiederum Verhaltensregeln festgelegt werden (etwa Community-Standards).[64] Das normative System wird dadurch noch komplexer, dass etwa Grund- und Menschenrechte nicht nur staatliche Stellen binden, sondern – jedenfalls nach deutschem Rechtsverständnis – indirekt auch Unternehmen anhalten, etwa die Meinungsfreiheit der Personen, die ihr Angebot nutzen, zu wahren. Zudem hat sich in der wissenschaftlichen Beobachtung die Erkenntnis durchgesetzt, dass die existierenden technischen Elemente ihrerseits eine normative Wirkung entfalten, die die Entwicklung prägt. Erst in der Zusammenschau all dieser Elemente zeigt sich der relevante Regelungsrahmen für das Verhältnis von Mensch und Maschine.

In dieser Stellungnahme befasst sich der Deutsche Ethikrat nicht im Detail mit dem hier kurz umrissenen reichhaltigen Fundus an ethischen Leitlinien und aktuellen wie noch in der Erarbeitung befindlichen rechtlichen Regulierungsansätzen, sondern knüpft auf einer anderen Ebene an diesen Rahmen an. Die in den folgenden zwei Kapiteln vorgelegte Analyse der Konsequenzen digitaler Entwicklungen für das menschliche Zusammenleben gründet auf einer philosophischen Auseinandersetzung mit den anthropologischen Grundbegriffen, die im Mittelpunkt des menschlichen Selbstverständnisses stehen (vgl. Kapitel 3), und entwickelt darauf aufbauend ein Verständnis von Mensch-Technik-Relationen, demzufolge es entscheidend darauf ankommt, wie die Delegation menschlicher Tätigkeiten an Maschinen und algorithmische Systeme auf zentrale anthropologische Konzepte zurückwirkt und dabei insbesondere menschliche Autorschaft erweitert oder vermindert (vgl. Kapitel 4). Im zweiten Teil der Stellungnahme wird dieses Konzept auf die vier ausgewählten Handlungsfelder Gesundheit, Bildung, Kommunikation und Verwaltung

64 Wielsch 2018; Kumkar 2022; Mast 2023.

angewendet und mündet jeweils in bereichsspezifische Empfehlungen, bevor im dritten Teil die sektorübergreifend identifizierten normativen Querschnittsthemen noch einmal näher beleuchtet werden.

3 ZENTRALE BEGRIFFE UND PHILOSOPHISCHE GRUNDLAGEN

3.1 Künstliche Intelligenz: begriffliche Analyse

Der Begriff der Künstlichen Intelligenz wurde und wird keineswegs immer einheitlich verwendet: Seine Bedeutung hat sich im Laufe der Jahre verändert und unterscheidet sich sowohl innerhalb als auch zwischen verschiedenen Berufsgruppen und Disziplinen. Eine Unterscheidung, welche in der KI-Forschung selbst, mehr noch aber in der Reflexion sowie der medialen und öffentlichen Debatte über KI eine große Rolle spielt, ist die Unterscheidung zwischen sogenannter *schwacher* und *starker* KI.[65] Während sich insbesondere die aktuelle Forschung innerhalb der Informatik vorrangig mit Fragen der schwachen KI beschäftigt, sind öffentliche Debatten ebenso wie manche Fachdiskurse regelmäßig geprägt von Visionen einer starken – also menschenähnlichen oder gar menschliche Fähigkeiten übertreffenden – KI, die mit Sorgen ebenso wie mit Hoffnungen behaftet ist. Um die hinter den jeweiligen Begriffsverwendungen stehenden substanziellen Annahmen über die Grundlagen menschlicher Intelligenz und KI herauszuarbeiten, ist eine Klärung der Begriffe notwendig. Dabei zeigt sich, dass unterschiedliche Einschätzungen hinsichtlich der Wahrscheinlichkeit und dem Erwünschtsein einer starken KI auch von unterschiedlichen Konzeptualisierungen des mit KI umschriebenen Phänomenbereichs abhängen.

Neben der heute häufig im Kontext der Reflexion auf KI vornehmlich verwendeten Unterscheidung von starker versus schwacher KI, sind in der Geschichte der KI und auch in der

65 Nida-Rümelin 2022a; 2022b.

gegenwärtigen internationalen Debatte andere Unterscheidungen in Gebrauch, um unterschiedliche Formen oder Grade der Annäherung von KI an menschliche Intelligenz zu beschreiben. Dies sind insbesondere die nachfolgend noch näher betrachteten Unterscheidungen von *spezieller* versus *allgemeiner/genereller* KI sowie *enger* versus *breiter* KI. Darüber hinaus werden in den letzten Jahren, vor allem im Kontext des Transhumanismus[66] sowie der Diskussionen um Singularität[67], die Begriffe Artificial General Intelligence oder Super-Intelligence verwendet, die im Folgenden allerdings keine Rolle spielen werden.

Den jeweiligen Unterscheidungen liegt trotz einzelner Bedeutungsunterschiede das Bemühen zugrunde, menschliche Intelligenz als Maßstab der Bestimmung von KI heranzuziehen. Über diesen Maßstab glaubt man zu verfügen, weil Intelligenztests eine (wenn auch in Grenzen) erfolgreiche Operationalisierung des Begriffs der Intelligenz erlauben (vgl. Abschnitt 3.2.1). Auf dieser Grundlage erscheint es möglich, menschliche Intelligenz zu simulieren bzw. das Vorliegen von Intelligenz in Maschinen identifizieren und messen zu können. Eine Simulation menschlicher Intelligenz stand lange im Zentrum der Forschung zu KI und der Reflexion über ihre Möglichkeiten und Grenzen. Auch wenn Methoden der KI für ganz andere Zwecke verwendet werden, beispielsweise die Nutzung zur Auswahl passender Werbung, so gilt seit der Veröffentlichung von Turings „Can machines think?" im Jahr 1950 und seiner Formulierung des berühmten Turing-Tests

66 Unter Transhumanismus werden Positionen verstanden, wonach die digitalen Technologien es uns erlauben, die Beschränktheiten der menschlichen Existenzform zu überwinden, und wir diese Möglichkeiten nutzen sollten, um alte Menschheitsträume zu verwirklichen, wie die einer durch Verbindungen von Gehirn und Computer (Brain Computer Interfaces) erzeugten Vervielfachung der Intelligenz oder der individuellen Fortexistenz in Form einer Softwarekopie des eigenen Gehirns.

67 Der Begriff der Singularität verweist auf einen möglichen Umschlagpunkt in der Zukunft, ab dem Künstliche Intelligenz menschliche Intelligenz in jeder Hinsicht übertreffen und sich fortan unkontrolliert selbst weiterentwickeln könnte.

(vgl. Abschnitt 2.1) die äußere Ununterscheidbarkeit zwischen menschlichen und maschinellen kognitiven und operativen Leistungen weithin als „Lackmustest" für das Vorliegen von KI.

Die Unterscheidung zwischen *spezieller* und *allgemeiner/ genereller KI* wurde in den 1950er-Jahren eingeführt, um die damals aktuelle Forschung zu KI von der Vision einer menschenähnlichen KI abzugrenzen. Während Maschinen zumeist für spezielle Tätigkeiten konstruiert sind, kennzeichnet menschliche Intelligenz eine Vielfalt von Kompetenzen und ein weites Spektrum unterschiedlicher Zwecksetzungen. Man erhoffte sich vom Computer von Anbeginn das Potenzial einer „universellen Maschine", die alle Aufgaben lösen könne, die sich binär, zum Beispiel als eine Abfolge von Nullen und Einsen, darstellen ließen.[68] Die Hoffnung war, dass der Computer eine generelle KI erreichen könnte, die über das Lösen spezifischer Aufgaben weit hinausgeht. Ein Ziel der frühen Forschung war die Entwicklung des Computers zum General Problem Solver. Dieser sollte basierend auf menschlichen Heuristiken Problemlösestrategien für ausreichend formalisierbare Probleme in verschiedenen Kontexten liefern. Er kann als Vorläufer heutiger Expertensysteme gelten. Die Komplexität der Aufgaben, die mit ihm gelöst werden konnten, blieb allerdings beschränkt. Die Vision einer allgemeinen KI jedoch hat bis heute Bestand.

Das Begriffspaar *speziell* und *allgemein* enthielt zumeist sowohl explizite Annahmen über die Breite des Fähigkeitenspektrums einer KI als auch Erwartungen zur grundlegenden Qualität der zukünftig daraus resultierenden KI (bis hin zum ontologischen Status). Entsprechend wurde im Verlauf weiter differenziert. Etabliert sind inzwischen die Unterscheidungen zwischen *enger* und *breiter* KI sowie *schwacher* und *starker* KI. Die Forschung zu KI ist heute zumeist auf einen klar umrissenen Anwendungsbereich begrenzt, etwa die Interpretation

68 Moor 1985.

von Röntgenaufnahmen, und somit ein Fall von enger KI. Doch auch hier besteht die Vision einer breiten Ausweitung des Spektrums. Dabei ist zu betonen, dass die Unterscheidung zwischen engen und breiten Formen der KI, ebenso wie die Begriffe *speziell* und *allgemein*, nicht als Gegensätze zu verstehen sind, sondern jeweils Endpunkte eines Fähigkeitenspektrums beschreiben.

Ein aktueller Bereich, in dem die Frage nach breiter KI verhandelt wird, umfasst Sprachverarbeitungssysteme wie das im vorherigen Kapitel bereits erwähnte GPT-3 und ChatGPT. Solche Systeme produzieren Texte, bei denen oft schwer oder unmöglich zu erkennen ist, ob sie von einem Menschen oder einer Maschine verfasst wurden. Erste prägnante Beispiele entfachten eine intensive philosophische und auch öffentliche Debatte dazu, ob diese Form der Textproduktion nur eine Verbreiterung des Fähigkeitenspektrums darstellt oder bereits als ein Übergang zur oder gar eine Manifestation einer generellen oder starken KI gelten könne – wenn auch nur mit Blick auf das Sprachvermögen.[69]

Die Verwendung der verschiedenen Begriffe zur Charakterisierung Künstlicher Intelligenz als spezifischer, enger oder schwacher KI einerseits sowie allgemeiner, breiter oder starker KI andererseits verweist nicht nur auf Differenzen zwischen den beiden jeweiligen Arten bzw. Polen von Intelligenz. Dahinter verstecken sich, insbesondere beim Begriffspaar der schwachen und starken KI, vielmehr auch unterschiedliche Verständnisse von Intelligenz sowie unterschiedliche

69 Der Titel eines Artikels des MIT Technology Review fasst diese Debatte mit Blick auf GPT-3 zusammen: „OpenAI's new language generator GPT-3 is shockingly good – and completely mindless" (https://www.technologyreview.com/2020/07/20/1005454/openai-machine-learning-language-generator-gpt-3-nlp [22.12.2022]). Dieser spielt auf die Kritik von John Searle an (vgl. Abschnitte 3.1 und 3.4.2). Mit Blick auf jüngere Chatbots löste 2022 zudem der ehemalige Google-Mitarbeiter Blake Lemoine eine Debatte aus, als er postulierte, der von Google entwickelte Chatbot LaMDA sei bei Bewusstsein – ein Vorstoß, der später zu seiner Entlassung führte (https://www.washingtonpost.com/technology/2022/06/11/google-ai-lamda-blake-lemoine [31.01.2023]).

Positionen hinsichtlich der Kernfrage, ob es qualitative und kategorische oder nur quantitative und prinzipiell überwindbare Unterschiede zwischen menschlicher Intelligenz und KI gibt.

Wichtig für die jeweilige Beantwortung dieser Frage ist zum einen die Differenz hinsichtlich der Breite bzw. Enge des Fähigkeitenspektrums der KI. Die meisten Anwendungen entfalten ihre jeweilige Leistung auf klar umrissenen, engen Gebieten oder Domänen wie beispielsweise dem Spielen von Schach oder Go. Hier sind sie im direkten Vergleich Menschen inzwischen klar überlegen. Sprachproduktionssysteme wiederum sind zwar ebenfalls auf einen Kompetenzbereich beschränkt (sprachliche Ein- und Ausgabe), jedoch mittlerweile nicht mehr auf eine Domäne, *über* die gesprochen wird; hier erfolgt also eine jedenfalls funktionelle Verbreiterung des Fähigkeitenspektrums. Dennoch fehlt auch ihnen jedwedes sprachliche Verständnis über die Bedeutung der rezipierten oder produzierten Worte, operieren sie doch allein auf Basis der Wahrscheinlichkeit von Wortkombinationen.

Zum anderen hatte der Philosoph John Searle schon 1980 gegen den Turing-Test die These aufgestellt, dass die bloße Ununterscheidbarkeit von menschlicher und maschineller Sprachperformanz nicht ausreicht, um ein Text*verständnis* anzunehmen.[70] Hier geht es also zusätzlich darum, ob es jenseits einer quantitativen Differenz auch einen kategorialen Unterschied zwischen Mensch und Maschine gibt, beziehungsweise darum, ob Intelligenz an bestimmte mentale Voraussetzungen geknüpft ist, welche über die bloße Simulation von Verständnis hinausgehen. Anders formuliert ergibt sich also die Frage, ob Intelligenz in allgemeiner oder starker Form jemals

70 Searle 1980. Searles Gedankenexperiment und seine Kritik werden ausführlicher im Abschnitt 3.4 behandelt.

vollumfänglich Maschinen zukommen kann oder ob dafür spezifisch menschliche Eigenschaften Voraussetzung sind.[71]

Die Differenzen hinter den Begriffspaaren schwache versus starke, enge versus breite und spezielle versus allgemeine KI lassen sich vor diesem Hintergrund wie folgt zusammenfassen:

Zum einen geht es um die *Breite der Fähigkeiten*, über die eine KI verfügt, sowohl graduell innerhalb von Domänen (z. B. Sprachverarbeitung) als auch bereichsübergreifend (z. B. Sprache und Motorik, Situationserfassung). Zum anderen geht es um die Antwort auf die Frage, ob die *Simulation von Intelligenz* mit *Intelligenz* gleichzusetzen ist oder ob es einen kategorischen Unterschied zwischen der *Simulation von Verständnis* und *genuinem* (z. B. sprachlichem) *Verständnis* gibt, der für „echte" Intelligenz essenziell ist, über die jedenfalls die in dieser Stellungnahme diskutierten Systeme nicht verfügen.[72]

Am Beispiel der Sprachproduktionssysteme lassen sich die vorgestellten unterschiedlichen Verständnisse von KI noch einmal veranschaulichen. Bereits die Antwort auf die Frage, ob sie Beispiele für breite bzw. allgemeine KI sind, hängt von den jeweiligen Annahmen ab: Zwar sind einzelne Sprachproduktionssysteme nicht auf eine Domäne beschränkt und haben somit ein durchaus breites Funktionenspektrum. Um allerdings die Anforderungen an breite bzw. allgemeine KI zu erfüllen, würde gemeinhin verlangt, dass das System nicht nur im Bereich der Sprache menschliche Kompetenz (nahezu) perfekt simuliert, sondern dies eben auch zeitgleich in (allen) anderen Bereichen vermag, die gemeinhin im menschlichen Kontext als intelligentes Verhalten klassifiziert werden, wie beispielsweise koordinierte Bewegung im Raum. Tatsächlich deuten bestimmte, auch bei sehr guten Sprachproduktionssystemen

71 Die Frage tierlicher Intelligenz wird im Rahmen dieser Stellungnahme ausgeklammert; vgl. dazu Huber 2021. Zur grundsätzlicheren Frage nach mentalen Zustände bei Tieren vgl. Deutscher Ethikrat 2020b.

72 Umstritten blieb, ob diese Aussage sich lediglich auf die aktuellen Systeme und die der absehbaren Zukunft bezieht oder grundsätzlich zu verstehen ist.

auftretende Fehlleistungen darauf hin, dass die hohe Leistungsfähigkeit nicht auf einem inhaltlichen Verständnis der Texte beruhen kann. Allgemein intelligentes Verhalten ist bei den gegenwärtigen Systemen schon funktional noch in weiter Ferne und bedürfte zudem auch noch des Einbaus in einen physischen humanoiden Roboter. In einem humanoiden Roboter mit perfekten Bewegungsfähigkeiten und einer menschenähnlichen Mimik und Gestik würden manche ein Beispiel breiter oder gar starker KI sehen, wenn er in der Lage wäre, alle menschlichen kognitiven Fähigkeiten perfekt zu simulieren. Andere würden hingegen bestreiten, dass damit eine Form starker KI vorliegt, da auch eine perfekte Simulation nicht garantiere, dass ein solcher humanoider Roboter mentale Zustände aufweist, über Einsichts- und Urteilsfähigkeit sowie über emotive Einstellungen wie Hoffnungen und Ängste verfügt.

Die unterschiedlichen Konzeptionen hinter den diversen Begrifflichkeiten zur KI gehen auch auf verschiedene grundlegende anthropologische Theoriemodelle zurück (vgl. Abschnitt 3.4). Aus behavioristischer Sicht ist die Unterscheidung zwischen Simulans und Simulandum nicht sinnvoll, da diese epistemisch nicht unterscheidbar seien. In anderen Konzeptionen (z. B. phänomenologische Positionen oder solche der intentionalistischen Semantik) jedoch werden mentale Zustände realistisch, das heißt als Merkmale der ontologischen Ausstattung der Welt interpretiert. In solchen Konzeptionen wird an einer kategorischen Unterscheidung zwischen Simulation und Realisierung festgehalten, auch wenn diese Differenz epistemisch nicht unmittelbar zugänglich ist. Entscheidend für den Unterschied zwischen menschlicher Intelligenz und KI ist demnach das Vorhandensein bestimmter mentaler Eigenschaften wie beispielsweise Verständnis oder Bewusstsein. In dieser Stellungnahme werden die Begriffe zur Charakterisierung der unterschiedlichen Formen Künstlicher Intelligenz in der unten stehenden Weise verwendet. Es wird hierbei vorausgesetzt, dass die Unterscheidung zwischen enger und breiter

KI quantitativer bzw. gradueller Natur ist, die Entstehung einer starken KI jedoch einen qualitativen Sprung bedeuten würde:

- *Enge KI:* KI-Anwendungen, die menschliche Fähigkeiten in einer Domäne simulieren bzw. Verfahren wie maschinelles Lernen verwenden, um spezifische Aufgaben zu erfüllen oder spezifische Probleme zu lösen. Nahezu alle derzeit verwendeten KI-Anwendungen fallen in diese Kategorie.
- *Breite KI:* Breite KI-Anwendungen erweitern das Spektrum ihrer Anwendbarkeit über einzelne Domänen hinaus. Sprachproduktionssysteme wie etwa GPT-3 können als Beispiele für breiter werdende KI gelten, da sie zwar nicht domänenspezifisch, jedoch weiterhin auf sprachliche Ein- und Ausgabe beschränkt sind. Eine mögliche Zukunftsvision breiter KI wären Systeme, die solche Sprachkompetenzen mit weiteren kognitiven oder motorischen Kompetenzen zusammenführen, etwa durch Einbau in weitentwickelte Roboter.
- *Starke KI:* Der Begriff der starken KI wird für die Vision einer Künstlichen Intelligenz verwendet, die jenseits der möglicherweise perfekten Simulation menschlicher Kognition auch über mentale Zustände, Einsichtsfähigkeit und Emotionen verfügen würde.

3.2 Intelligenz und Vernunft

3.2.1 Intelligenz

Die im vorigen Abschnitt vorgestellten unterschiedlichen Deutungen von KI werden seit mindestens den 1970er-Jahren von kontroversen Diskussionen der Frage begleitet, was Computer können und nicht können bzw. demnächst können und nicht können werden.[73] Um diese Frage zu beantworten,

73 Dreyfus 1992.

müsste zunächst geklärt werden, von welchen Vorstellungen hinsichtlich der menschlichen Intelligenz dabei ausgegangen wird. Dazu wird jedoch in den sich damit beschäftigenden Wissenschaften, insbesondere der Psychologie, Philosophie und Informatik, keine einheitliche Antwort angeboten.

Aus psychologischer Perspektive ist Intelligenz als ein hypothetisches Konstrukt aufzufassen, das als solches zwar verbal umschrieben werden kann, zum Beispiel im Sinne von Verstehen, Urteilen und Schlussfolgern[74] oder zielgerichtetem Handeln, rationalem Denken und effektiver Auseinandersetzung mit der Umwelt[75], aber nicht beobachtbar ist, sondern anhand von Indikatoren in relevanten Aspekten operationalisiert werden muss. In diesem Sinne sind Intelligenztests als Situationen aufzufassen, in denen Menschen Verhalten zeigen können, das vor dem Hintergrund eines theoretischen Vorverständnisses oder einer zugrunde gelegten Definition als mehr oder weniger „intelligent" bezeichnet werden kann. Dabei können die gewählten Operationalisierungen – auf der Ebene von Teilkomponenten und Subskalen wie auf der Ebene von Testitems – sehr unterschiedlich sein. Sie bewähren sich im Kontext der Bestimmung von Gütekriterien, auf deren Grundlage abgeschätzt werden kann, inwieweit die Durchführung, Auswertung und Interpretation der Testung als objektiv, verlässlich (reliabel) und valide angesehen werden können.

Intelligenz wird hier im Sinne eines Abweichungsquotienten verstanden; der Mittelwert des (normalverteilten) Merkmals liegt in der Grundgesamtheit per definitionem bei 100, die Standardabweichung bei 15.[76] Dies hat Auswirkungen auf die Testkonstruktion, denn bei der Auswahl potenzieller Testitems muss berücksichtigt (geschätzt) werden, wie sich diese in der Grundgesamtheit verteilen, mit anderen Items korrelieren und zwischen verschiedenen Fähigkeitsniveaus unterscheiden.

74 Binet/Simon 1905.
75 Wechsler 1944.
76 Das bedeutet, dass ca. 68 Prozent der Menschen einen IQ zwischen 85 und 115 sowie ca. 95 Prozent einen IQ zwischen 70 und 130 aufweisen.

Bei der Validierung von Intelligenztests interessieren Korrelationen mit anderen Verfahren und Merkmalen, die aus theoretischer Perspektive als Indikatoren der Ausprägung verwandter und nicht verwandter Konstrukte betrachtet werden können und deshalb eine bestimmte Höhe aufweisen bzw. diese nicht überschreiten sollten. Darüber hinaus ist die Möglichkeit, auf der Grundlage aktueller Messwerte zukünftige Leistungen bzw. interindividuelle Unterschiede in relevanten Merkmalen (z. B. Erfolg versus Misserfolg in Schule oder Beruf) zu prognostizieren, von großem Interesse.

Die Wechsler Adult Intelligence Scale, auf der die bekanntesten modernen Intelligenztests aufbauen, wurde 1955 von David Wechsler vorgestellt und bereits 1956 in einer deutschsprachigen Version, dem Hamburg-Wechsler-Intelligenztest für Erwachsene, veröffentlicht. Die aktuelle Version des Tests besteht aus 15 Untertests (zehn Kerntests und fünf optionale Untertests), welche unterschiedliche kognitive Fähigkeiten prüfen, die in vier Indizes zusammengefasst werden: Sprachverständnis, wahrnehmungsgebundenes logisches Denken, Arbeitsgedächtnis und Verarbeitungsgeschwindigkeit.[77] Hierbei ist wichtig festzuhalten, dass es keine Aufgaben gibt, die diese Dimensionen direkt messen. Vielmehr sind die verwendeten Skalen induktiv aus den im Kontext der Testkonstruktion und -normierung empirisch ermittelten Korrelationen zwischen Unterskalen und Außenkriterien abgeleitet. Das heißt, die Struktur der Intelligenz ergibt sich wesentlich induktiv aus der empirischen Erfassung und Validierung unterschiedlicher, aus dem theoretischen Vorverständnis der Testautorschaft abgeleiteten Aspekte bzw. Dimensionen kognitiver Leistungsfähigkeit (z. B. der Orientierung an einem Generalfaktormodell in der Tradition von Charles Spearman versus einem Modell voneinander unabhängiger Primärfaktoren in der Tradition von Louis Leon Thurstone).[78]

77 Petermann 2012, 21.
78 Sternberg 2020.

Die Frage, ob Intelligenz eine einheitliche Fähigkeit ist oder viele Fähigkeiten umfasst, die gegebenenfalls auch voneinander unabhängig sein können, ist empirisch nicht eindeutig zu klären – die dimensionale Struktur wird auf der Grundlage von vorab festgelegten Modellannahmen bzw. Restriktionen ermittelt, die ihrerseits nicht Gegenstand einer empirischen Überprüfung werden können. Allerdings lässt sich festhalten, dass empirisch durchweg positive Korrelationen zwischen den verschiedenen Untertests bzw. Indizes nachzuweisen sind. Im erwähnten Wechsler-Intelligenztest bedeutet dies etwa, dass höhere Werte in den Untertests des Index Sprachverständnis statistisch mit höheren Werten in den anderen drei Indizes einhergehen, auch wenn das Ausmaß der Korrelationen zwischen den verschiedenen Untertests je nach Nähe der Aufgaben variiert.[79] Diese durchweg positive Korrelation führte zur Annahme des sogenannten Generalfaktors der Intelligenz, welcher den Anteil der allgemeinen Intelligenz bzw. der kognitiven Leistungen zugrunde liegenden allgemeinen geistigen Fähigkeit bezeichnet, deren Ausprägung sich – weil unidimensional – in einem einzigen Wert ausdrücken lässt, und ca. 40 Prozent der in Leistungsmessungen beobachteten Varianz erklärt. Die übrigen 60 Prozent lassen sich demnach auf unterschiedliche spezifische Fähigkeiten zurückführen.

Innerhalb der Allgemeinen Psychologie und insbesondere mit der Entwicklung der Kognitionspsychologie in den 1970er-Jahren rückte zunehmend die Analyse der Prozesse in den Fokus, die nötig sind, um die Aufgaben der Intelligenztests zu lösen. Umfangreiche Forschungen zur Bearbeitung informationsverarbeitender Aufgaben wie Informationscodierung und geteiltes Hören führten zu der Annahme, dass (verbale) Intelligenz durch die Fähigkeit zur Auswahl und Benutzung

79 Die durchschnittliche Korrelation zwischen allen Untertests liegt bei 0,45. Aufgrund ihres statistischen Ursprungs sind die Korrelationen innerhalb der vier Indizes höher als zwischen diesen. Die höchste Korrelation von 0,74 besteht zwischen dem Wortschatz-Test und dem Test für allgemeines Verständnis (Deary 2020, 9).

von Informationsverarbeitungsmethoden bestimmt wird.[80] Die Kernthemen der kognitionspsychologischen Forschung jedoch, nämlich Denken, Problemlösung und Entscheidungsfindung, die auch außerhalb der Disziplin der Psychologie häufig mit dem Begriff „Intelligenz" assoziiert werden und auf die häufig in der Entwicklung von KI Bezug genommen wird, fallen in der Psychologie nicht unter den Intelligenzbegriff. Dies mag mit der zunehmenden Spezialisierung innerhalb der Psychologie zusammenhängen. So ist die Erfassung menschlicher Intelligenz ein Teilgebiet der Differentiellen Psychologie, die sich mit Unterschieden zwischen Menschen befasst, wohingegen die Allgemeine Psychologie sich mit den Grundlagen von Wahrnehmung, Lernen sowie insgesamt menschlicher Kognition und Emotion beschäftigt.

Verwiesen sei in diesem Zusammenhang auch auf die Unterscheidung zwischen Intelligenz und Kreativität[81], wobei Letztere in Anlehnung an Joy Paul Guilford, den Begründer der modernen Kreativitätsforschung, als flüssige, flexible und ursprüngliche Erzeugung von Konzepten von Lösungen für neuartige Probleme definiert werden kann.[82] Von Interesse ist hier seine Unterscheidung zwischen konvergentem und divergentem Denken.[83] Im Unterschied zum konvergenten Denken, das durch logische Schlussfolgerungen zu einer einzigen oder besten Lösung gelangt (wobei das Ergebnis mehr oder weniger vollständig durch die vorhandene Information determiniert ist), liefert das für Kreativität charakteristische divergente Denken mehrere alternative Lösungen, die jeweils den gegebenen Anforderungen entsprechen. Dabei gelten die Anzahl der generierten Lösungen und deren Qualität als Maß für die Ausprägung des divergenten Denkens. Neben dem divergenten Denken wurden und werden auch weitere kognitive Prozesse als zentrale Voraussetzungen für Kreativität diskutiert. Zu

80 Hunt/Lunneborg/Lewis 1975.
81 Kruse/Schmitt 2011; Lubart 2018.
82 Guilford 1950.
83 de Vries/Lubart 2017.

nennen sind hier insbesondere Fähigkeiten und Fertigkeiten im Bereich von Wahrnehmung, Problemdefinition, Einsicht, Induktion, Bildung von Analogien und ungewöhnlichen Assoziationen, die Bewertung von Ideen und die Organisation von Wissenssystemen.

Während der Begriff der Intelligenz ursprünglich auf ein Ensemble menschlicher Leistungen verweist, wie sie in klassischen Intelligenztests gemessen werden[84], hat sich der Blick auf Intelligenz in jüngerer Zeit sukzessive erweitert. So entstanden Konzepte wie die der sozialen bzw. emotionalen Intelligenz, welche einerseits den Begriff der Intelligenz auf weitere Fähigkeiten ausweiteten sowie andererseits die Wechselwirkungen zwischen Emotion und Kognition sowie den sozialen und kulturellen Aspekt von Intelligenz in den Blickpunkt rückten. Darüber hinaus entwickelte sich rund um die Stichwörter *embodied*, *embedded*, *enactive* und *extended* Kognition ein Forschungsfeld, das in Philosophie, Psychologie und Robotik die Rolle des Körpers einerseits und der Umwelt andererseits für Intelligenz und kognitive Leistungen erforscht.[85]

Spätestens diese Erweiterungen werfen die Frage auf, wie die Übertragung des Intelligenzbegriffs auf technische Artefakte zu verstehen ist. Bei einigen Merkmalen wie beispielsweise der elementaren Rechenfähigkeit, dem logischen Schlussfolgern oder Gedächtnisleistungen entsteht der Eindruck, dass menschliche Fähigkeiten eindeutig auf technische Artefakte übertragen werden können. Auch diesbezüglich sind schon kritische Fragen zu stellen, zum Beispiel ob die menschliche Erinnerungsfähigkeit in gleicher Weise eine Gedächtnisleistung ist wie die Aktivierung eines technischen Speichers.

---

84 Verwiesen sei hier auf die von Louis Leon Thurstone faktorenanalytisch ermittelten sieben Primärfaktoren induktives Schließen, räumliches Vorstellungsvermögen, Wahrnehmungsgeschwindigkeit, Rechenfähigkeit, verbales Verständnis, assoziatives Gedächtnis und Wortflüssigkeit (Thurstone 1938; Thurstone/Thurstone 1941) sowie die Differenzierungen zwischen fluider versus kristalliner Intelligenz (Horn/Cattell 1966) und kognitiver Mechanik versus Pragmatik (Baltes 1987).

85 Clark 2012.

Einerseits sind die quantitativen Leistungen technischer Speicher der menschlichen Erinnerungsfähigkeit um Größenordnungen überlegen. Andererseits sortiert der Mensch seine Gedächtnisleistungen beispielsweise nach der jeweils kontextuell bestimmten Bedeutung, während ein technischer Speicher unterschiedslos je nach den technischen Vorgaben Daten aufnimmt oder nicht. Bei Intelligenzleistungen mit emotiven und kreativen Qualitäten verstärkt sich der Verdacht, dass es sich hierbei um anthropomorphe Übertragungen handelt. Die Klärung solcher Vergleichbarkeitsprobleme hängt somit wesentlich von den Kriterien ab, durch die man eine spezifisch menschliche Intelligenzleistung bestimmt sieht. Man sollte daher die Verwendung des Ausdrucks „Intelligenz" in der Wortverbindung „Künstliche Intelligenz" eher als eine Metapher einordnen, deren Beschreibungs- und Erklärungsfunktion genauerer Aufklärung bedarf.

3.2.2 Vernunft

Bereits lange vor der Einführung des Begriffs der Intelligenz wurde der Begriff der Vernunft verwendet, um die spezifische menschliche Fähigkeit zu kennzeichnen, sich in der Welt zu orientieren, selbstverantwortlich zu handeln und so der eigenen Lebenspraxis eine kohärente Struktur zu geben. Intelligenz ist für Vernunft eine wichtige Voraussetzung, aber keine hinreichende Bedingung. Der Begriff der Vernunft gehört zu den basalen Grundkategorien menschlicher Selbst- und Weltdeutung, die unsere Kultur seit der Antike maßgeblich geprägt haben. Schon das weite Wortfeld (griechisch: *logos, nous, dianoia, phronesis*; lateinisch: *ratio, mens, intellectus, prudentia*) deutet darauf hin, dass es sich um einen überaus komplexen Begriff handelt, der vielfältige Binnendifferenzierungen kennt und verschiedene (kognitive) Teilkompetenzen umfasst. Strukturell geht es um ein mehrdimensionales Beziehungsgefüge von Denk-, Reflexions- und Operationsformen,

das in seiner Gesamtheit im Dienste einer möglichst adäquaten Wirklichkeitserschließung steht und in einen komplexen sozialen und kulturellen Kontext verwoben ist. Als Inbegriff bestimmter Ansprüche, denen wir uns im Denken, Sprechen, Erleben und Handeln unterstellen, umfasst der Vernunftbegriff unterschiedliche – propositionale und nicht propositionale – Wissensformen und Rationalitätstypen, die von methodisch-prozeduralem Know-how über ästhetische Wahrnehmungsfähigkeit und Kreativität sowie verschiedene soziale Interaktionsfähigkeiten bis hin zu einer umfassenden Lebensführungskompetenz reichen. Von grundlegender Bedeutung für die hier behandelte Thematik ist dabei die Gegenüberstellung von *theoretischer Vernunft*, die sich auf den Erkenntnisgewinn richtet, um zu wahren empirischen oder apriorischen Urteilen zu gelangen, und *praktischer Vernunft*, die auf ein kohärentes, verantwortliches Handeln abzielt, um ein gutes Leben zu ermöglichen.

Vor allem im Blick auf den Gebrauch der *theoretischen* Vernunft, der primär auf Erkenntnisgewinn durch die Formulierung wahrer empirischer Urteile abzielt, scheinen sich zumindest prima facie einige Parallelen zur Arbeitsweise von KI-Systemen aufzudrängen. So spielen in beiden Bereichen Fähigkeiten der Informationsverarbeitung, des Lernens, des logischen Schlussfolgerns und konsistenten Regelfolgens sowie der sinnvollen Verknüpfung gespeicherter Daten eine zentrale Rolle. Bei näherer Betrachtung zeigen sich jedoch insofern gravierende Differenzen, als sich nicht nur die Arbeitsweise des menschlichen Gedächtnisses in mehrfacher Hinsicht vom technischen Speicher eines Computers unterscheidet, sondern auch die menschliche Urteilspraxis technisch nicht substituierbar ist. Auch wenn in diesem Zusammenhang die wahrheitstheoretischen Implikationen der Formulierung und Begründung deskriptiver Urteile nicht näher entfaltet werden können[86], ist doch darauf hinzuweisen, dass zumindest die

86 Bovens/Hartmann 2003.

bislang verfügbaren KI-Systeme die dafür relevanten Fähigkeiten des Sinnverstehens, der Intentionalität und der Referenz auf eine außersprachliche Wirklichkeit nicht besitzen.

Dieser Befund bestätigt sich auch bezüglich der im Rahmen dieser Stellungnahme besonders interessierenden *praktischen* Vernunft, die insofern noch weit komplexerer Natur ist, als ihr Ziel nicht nur in wohlbegründeten praktischen Einzelurteilen, sondern in einem möglichst richtigen und verantwortlichen Handeln besteht, das über einen langen Zeitraum aufrechterhalten wird, eine kohärente Ordnung der Praxis garantiert und damit ein insgesamt gutes Leben ermöglicht.[87] Dazu bedarf es mehrerer Einzelkompetenzen, deren Simulationsmöglichkeiten durch technische Artefakte gegenwärtig mit Blick auf die unterschiedlichen Relationen und Wechselwirkungen zwischen Mensch und Maschine (vgl. Abschnitt 4.3) kontrovers diskutiert werden. Ohne Anspruch auf Vollständigkeit seien dabei die folgenden acht Teilfähigkeiten exemplarisch besonders hervorgehoben:

Erstens braucht es ein *Verständnis* der für unsere Moralsprache konstitutiven evaluativen und deontischen Prädikatoren: Von einem vernünftigen Wesen erwarten wir, dass es über die Fähigkeit verfügt, die Bedeutung der verschiedenen, phänomenologisch gehaltvollen Ausdrücke zur Bezeichnung moralisch relevanter Güter, Werte und Haltungen sowie deontischer Prädikate zur Qualifizierung von Handlungen (wie richtig bzw. falsch) angemessen zu verstehen und situationsadäquat zu gebrauchen.

Zweitens wird ein *Unterscheidungs- und Einfühlungsvermögen* benötigt, um die moralisch relevanten Differenzen zwischen einzelnen moralischen Gütern, Werten, Handlungstypen und Lebensformen möglichst präzise und realitätsnah erfassen sowie anderen Menschen empathisch begegnen zu können.

87 Bormann 2021, 48–51.

Drittens muss die Fähigkeit zur *Abwägung* konfligierender Güter und Werte vorliegen: Die praktische Vernunft beinhaltet auch ein deliberatives Vermögen, das immer dann ins Spiel kommt, wenn komplexe Handlungsstrategien entwickelt werden müssen oder mehrere moralisch bedeutsame Gesichtspunkte aufgrund bestimmter ungünstiger Umstände in einer konflikthaften Beziehung zueinanderstehen. Mittels des Vermögens der Güterabwägung vermag die handelnde Person nicht nur zu erkennen, welche Güter in zeitlicher Hinsicht prioritär erstrebt oder gesichert werden müssen, um bestimmte Ziele zu erreichen, sondern welchen Gütern im Konfliktfall der Vorrang zuzuerkennen ist, um ein situativ richtiges Handeln zu ermöglichen.

Viertens bedarf es der Befähigung zum *reflektierten Umgang mit Regeln* unterschiedlicher Reichweite: Die praktische Vernunft schließt auch die Fähigkeit ein, moralische Regeln wie zum Beispiel Normen und Prinzipien verstehen, korrekt anwenden und falls nötig auch weiterentwickeln zu können, um Probleme zu lösen und ein realitätsadäquates kohärentes Handeln über längere Zeiträume zu ermöglichen. Obwohl ein Großteil des menschlichen Handelns von Routinen und Konventionen bestimmt wird, gibt es auch vielfältige Herausforderungen und Konfliktsituationen, die durch ein konventionelles oder gar starr deterministisches Regelfolgen allein gerade nicht zu bewältigen sind, sondern Kreativität und einen flexibleren Umgang mit regulatorischen Vorgaben auf der Grundlage eines unvertretbaren Aktes der praktischen Urteilskraft erfordern.[88]

88 In diesem Zusammenhang ist auf einen grundsätzlichen Unterschied zwischen dem algorithmischen gegenüber einem heuristischen Regelverständnis hinzuweisen. In einem algorithmischen Verfahren bilden die Regeln den Auswahlfilter, durch den die Fälle als Kandidaten überprüft und verworfen oder zur Überprüfung im nächsten Schritt angenommen werden. In algorithmischen Verfahren steht damit das Verhältnis von Regel und Fall fest. In einem heuristischen Verfahren werden die Regeln durch die Subsumtion eines Falles pragmatisch und semantisch mit-konstituiert, sodass nicht prä-prozedural feststeht, unter welche Regel der Fall gehört.

Fünftens wird die Fähigkeit zum *intuitiven Erfassen komplexer Handlungssituationen und Umstände* benötigt: Menschen müssen oft unter großem Zeitdruck weitreichende Entscheidungen treffen und dabei vielfältige Merkmale eines Handlungskontextes berücksichtigen.

Sechstens bedarf es eines *Urteilsvermögens*, mittels dessen Personen in der Lage sind, Entscheidungen zwischen verschiedenen Handlungsalternativen zu treffen und singuläre Handlungskonstellationen bestimmten generellen Handlungstypen zuzuordnen.[89]

Siebtens braucht es die Fähigkeit zur *Begründung* der eigenen moralischen Urteile und der ihnen korrespondierenden Praxis: Die Fähigkeit, Gründe zu geben und zu nehmen (*give and take reasons*) und sich im Urteilen und Handeln daran auszurichten, schließt neben der Bereitschaft zur kritischen Reflexion eigener partikularer Interessen auch die Fähigkeit ein, einen moralischen Standpunkt (*moral point of view*) einzunehmen, also die für die moralische Qualifikation einer Handlung relevanten Gründe aus der Dritte-Person-Perspektive zu beurteilen.

Achtens muss die Fähigkeit zur *Affekt- und Impulskontrolle* vorliegen, um die jeweils gefällten praktischen Urteile auch handlungswirksam werden zu lassen. Gerade bei der Verfolgung anspruchsvoller Ziele, die vielfältige Vorarbeiten und einen langen Atem verlangen, ist es wichtig, die erforderliche Willensstärke aufzubringen und zumindest solchen spontanen Affekten, Neigungen und Impulsen zu widerstehen, die den langfristigen Erfolg der jeweiligen Bemühungen gefährden oder sogar verunmöglichen können.

---

89 In Anspielung auf die durch das richterliche Judiz zu leistende intellektuelle Aufgabe hat Immanuel Kant für heuristische Verfahren dieser Art den Begriff der Urteilskraft geprägt. Das algorithmische Verfahren ordnet Kant (1968a) der bestimmenden Urteilskraft zu, der Domäne des Verstandes (KrV, B 360 f.). Das in den praktischen Disziplinen wie Pädagogik, Jurisprudenz oder Ökonomik zugrunde zu legende Verfahren der Suche nach der angemessenen Passung von Regel und Fall zeichnet Kant (1968b) als reflektierende Urteilskraft aus, die zur Domäne der praktischen Vernunft gehört (KU, AA V, 385).

Die Unterscheidung der verschiedenen Teilkompetenzen der Vernunft ist für die hier behandelte Thematik aus zwei Gründen bedeutsam: Erstens ist es durchaus möglich, dass es partielle Überschneidungen des Kompetenzprofils moderner KI-Systeme mit dem komplexen Phänomen menschlicher Vernunft gibt, was insbesondere im Bereich des Regelfolgens und der Weiterentwicklung vorgegebener Algorithmen der Fall sein dürfte. Zweitens ist zu berücksichtigen, dass die hier genannten, für den Bündelbegriff der „praktischen Vernunft" konstitutiven Fähigkeiten nicht einfach im Sinne isolierter Einzelelemente beziehungslos nebeneinanderstehen. Vielmehr ist von vielfältigen Wechselwirkungen, Rückkopplungen und Bedingungsverhältnissen zwischen ihnen auszugehen. Sie bilden einen integralen Bestandteil einer komplexen menschlichen Natur, die im Sinne einer leibseelischen Einheit zu verstehen ist. Menschliche Vernunft ist stets als *verleiblichte Vernunft* zu begreifen (vgl. Abschnitt 3.4.3). Nur so ist zu erklären, dass praktische Überlegungen überhaupt handlungswirksam werden können. Zurückzuweisen ist eine Deutung, die versucht, vernünftige Vollzüge aus einer rein individualistischen Perspektive zu rekonstruieren. Da jeder Mensch Teil einer sozialen Mitwelt und kulturellen Umgebung ist, die sich nachhaltig auf seine Sozialisation auswirken, müssen auch überindividuelle kulturelle Faktoren in die Deutung der praktischen Vernunft einbezogen werden. Ein angemessenes Verständnis insbesondere des praktischen Vernunftgebrauchs ist eng mit unserem basalen Selbstverständnis als handlungsfähige Personen verbunden. Da technische Artefakte in immer neuen Formen in die Handlungswelt der Menschen integriert werden, mit Menschen interagieren oder sogar Teilfunktionen menschlichen Handelns übernehmen, ist es wichtig, zunächst den Handlungs- und Verantwortungsbegriff zu klären.

3.3 Handlung und Verantwortung

3.3.1 Handlung

Auch wenn im Alltag gelegentlich alle möglichen Ereignisse als Handlungen bezeichnet werden, fassen die Normwissenschaften Ethik und Jurisprudenz und auch die Psychologie[90] den Handlungsbegriff oft enger. Dabei wird angenommen, dass Menschen in der Lage sind, aktiv, zweckgerichtet und kontrolliert auf die Umwelt einzuwirken und dadurch Veränderungen zu verursachen. Das bedeutet, dass nicht jedes menschliche Tun, das auf die Umwelt einwirkt, als Handlung zu verstehen ist, sondern nur solches, das zweckgerichtet, beabsichtigt und kontrolliert ist.[91] Unterstellt man, dass Maschinen nicht zweckgerichtet operieren, also keine Absichten haben, dann ist die Zuschreibung von Handlungen in Bezug auf Maschinen in diesem engen Sinne nicht möglich.

Für den Menschen, der im engen Sinn handelt, hat sich auf dem Hintergrund einer lang zurückreichenden Begriffsgeschichte[92] im Rahmen einer umfassenden Theorie praktischer Vernunft der Begriff der *Person* eingebürgert.[93] Personen sind Akteure, die Verantwortung für ihr Verhalten tragen, die die kognitiven Bedingungen erfüllen, um Handlungsoptionen zu erkennen, und die Gründe deliberieren können, also über ein

90 Vgl. etwa die verschiedenen (etablierten) Handlungsphasenmodelle.

91 Hier wird eine Form der Handlungserklärung unterstellt, die die Warumfrage (im Sinne von „Warum hast du das getan?“) durch Angabe der Absicht bzw. Zwecke der Handlung beantwortet sieht (intentionale, teleologische Handlungserklärung, „Intentionalismus“). Ihr steht eine Form der Handlungserklärung gegenüber, die die Angabe der die Handlung auslösenden Ursachen als Antwort vorsieht (kausale Handlungserklärung, „Kausalismus“). Ob eine Handlung intentionalistisch oder kausalistisch erklärt werden sollte, ist keine Frage der Wahrheit/Falschheit der Erklärung, sondern eine Frage der Kontextadäquatheit der Handlungsdeutung. Für die Normwissenschaften steht die Frage der Absicht bzw. der Zweck der Handlung im Vordergrund der Betrachtung, ohne dass die Möglichkeit einer kausalen Handlungserklärung in Abrede gestellt wird (Horn/Löhrer 2010).

92 Fuhrmann 1989.

93 Quante 2007.

hinreichendes Maß theoretischer und praktischer Vernunft verfügen.

Der Handlungsbegriff ist auch deswegen so bedeutsam in der Diskussion um maschinelle Fertigkeiten und KI, da dieser Diskurs sich etwa seit der Jahrtausendwende von der Konzentration auf möglicherweise kognitive Kompetenzen abgewandt hat und inzwischen zunehmend auf praktische Kompetenzen konzentriert.[94] Nicht in welcher Weise, wenn überhaupt, Maschinen denken, sondern in welchem Sinne Maschinen handeln können, steht verstärkt im Vordergrund. Diese Verschiebung der Aufmerksamkeit geht Hand in Hand mit technischen Entwicklungen hin zu maschinellen Systemen, die nicht lediglich auf hohe Informationsverarbeitungskapazität setzen, um menschliches Handeln zu unterstützen, sondern jenes teilweise gar ersetzen können oder sollen.

Automatisierte oder algorithmische Entscheidungssysteme (Automated/Algorithmic Decision Making Systems, ADM-Systeme) zum Beispiel erstellen auf Basis von Berechnungen Prognosen darüber, wie geeignet eine sich bewerbende Person für eine Stelle ist oder mit welcher Wahrscheinlichkeit Menschen Kredite zurückzahlen oder straffällig werden (vgl. Kapitel 8). Auch wenn diese Systeme oftmals nur der Unterstützung der menschlichen Entscheidung dienen, so können Entscheidungen auch komplett an jene delegiert werden. Damit stellt sich die Frage, in welchem Sinne solche maschinellen Vollzüge außerhalb des obigen engen Handlungsbegriffs doch in bestimmten Kontexten als Handlungen in einem weiteren Sinne wahrgenommen werden können oder berücksichtigt werden müssen. Daran anknüpfend gibt es einen Diskurs, ob und inwieweit zunehmend eigenständige, das heißt ohne menschliches Zutun funktionierende maschinelle Systeme als „Agenten" in der Folge für ihr „Handeln" verantwortlich

94 Dreyfus 1992.

gemacht werden können, etwa mit Blick auf Fragen der Haftung (vgl. Abschnitt 2.2.5).[95]

Den folgenden Überlegungen liegt mit Blick auf die oft schillernde Verwendung zentraler ethisch relevanter Begriffe wie *Handlung* oder *Verantwortung* im Kontext der zeitgenössischen KI-Debatte die Annahme zugrunde, dass wenigstens drei Klassen von Entitäten terminologisch klar gegeneinander abzugrenzen sind. Dies wären erstens Pflanzen und Tiere, die zwar in vielfältiger Weise auf ihre Umwelt reagieren können, in ihrem jeweiligen Repertoire aber doch so begrenzt sind, dass der ausgeführte enge Handlungsbegriff auf sie nicht anwendbar ist.[96] Zweitens sind Menschen in dem Maße im engen Sinn als handlungsfähig zu bezeichnen, als sie dazu imstande sind, absichtlich Veränderungen zu bewirken, wobei solche Handlungen nicht nur Taten, sondern auch deren bewusstes und absichtliches Unterlassen umfassen können.[97] Drittens gibt es technische Artefakte unterschiedlicher Komplexitätsgrade, deren jeweilige Vollzüge oder Operationen zwar Veränderungen in der Welt bewirken und flexibel mit anspruchsvollen Herausforderungen der menschlichen Lebenswelt umgehen können.[98] Da sie diese Veränderungen aber nicht absichtlich herbeiführen, haben sie selbige daher auch nicht in einem moralischen und rechtlichen Sinne zu verantworten (vgl. Abschnitt 3.3.2).

Auch wenn es zwischen diesen drei Klassen von Entitäten unterschiedliche Arten von Wechselwirkungen (vgl. Abschnitt 4.3) geben kann, scheint es sinnvoll, den Handlungsbegriff im

95 Hilgendorf 2014; 2020; Ebers et al. 2020.

96 Ausnahmen werden für einige hochentwickelte Tiere diskutiert, darunter zum Beispiel Primaten und Rabenvögel (Huber 2021).

97 Für Unterlassungen, nicht aber für Nichthandeln, kann man getadelt oder verurteilt werden. Das Unterlassen bezieht sich demgemäß auf ein bestimmtes Handlungsschema (z. B. des Helfens) und ist nicht nur die unbestimmte Negation der Ausführung eines Handlungsschemas. Unterlassungshandlungen sind somit „Ereignisse in der Welt" und als solche ohne Weiteres mögliche Ursachen von Wirkungen (Folgen) in der Welt (Roxin/Greco 2020, 335 ff.; Bottek 2014).

98 Rammert/Schulz-Schaeffer 2002.

engen Sinne Menschen vorzubehalten, um inflationären Ausweitungen des Akteursstatus zu vermeiden und konzeptionelle Grenzziehungen zu ermöglichen. Entscheidend ist demnach das Konzept der Handlungsurheberschaft bzw. Autorschaft, das auf die universelle menschliche Handlungserfahrung verweist, sich selbst und andere im Hinblick auf bestimmte Ereignisse und Zustände als Urheber anzusehen.[99] Die Fähigkeit zur Handlungsurheberschaft kann als Grundlage von Autonomie betrachtet werden, also dafür, dass handelnde Menschen ihre Handlungen nach Maximen ausrichten können, die sie sich selbst setzen. Diese Konzeption schließt jedoch nicht aus, dass Handlungen mitunter auch durch Befolgen von Autoritäten und Traditionen ausgeführt werden. Wo dies geschieht, setzt das Konzept der Handlungsurheberschaft jedoch voraus, dass Menschen ihr eigenes Dasein in ein Verhältnis zu solchen Bestimmungen setzen können, etwa durch Überwindung, Widerstand oder Nachgeben.

Aus einer Handlung können neben den beabsichtigten Folgen auch nicht beabsichtigte, aber der handelnden Person erkennbare Folgen erwachsen. Auf diese Weise erscheint es möglich, auch fahrlässiges Tun zu erfassen, das im Kontext von KI eine große Rolle spielt.[100] Des Weiteren finden Handlungen umgeben von anderen raumzeitlichen Ereignissen statt, die als Umstände der Handlung für deren moralische und rechtliche Bewertung von Bedeutung sein können. Auch werden Handlungen zwar methodisch primär individuellen Akteuren zugesprochen; das schließt aber kollektive Handlungen nicht aus, bei denen die einzelnen Personen von vornherein in einem Kontext der Koordination agieren, ihre Handlungen auf

99 Deutscher Ethikrat 2017, 175 ff.; Nida-Rümelin 2020a, 376–408.

100 Ein Beispiel liefert etwa der Bereich des autonomen Fahrens: Unterläuft dem Programmierer ein Fehler, der später zur Schädigung von Verkehrsteilnehmern führt, und war dies für ihn vorhersehbar und vermeidbar sowie aus Gründen des überwiegenden Schutzes der Interessen der anderen Personen auch von Rechts wegen zu vermeiden, liegt hierin ein fahrlässiges Fehlverhalten, für das er zur Verantwortung zu ziehen ist (Hevelke/Nida-Rümelin 2015).

Kooperation bezogen sind und durch Kommunikation gestützt werden.

Der hier verwendete, eng gefasste Handlungsbegriff schließt nicht aus, dass Technologie erheblichen Einfluss auf menschliches Handeln oder die menschliche Handlungserfahrung haben kann. Technik beeinflusst und verändert Gesellschaft; gleichzeitig beeinflusst Gesellschaft die Technikentwicklung und den Technikeinsatz. Gerade die in den letzten Jahren stark zunehmende Durchdringung der menschlichen Lebenswelt mit informationstechnisch immer leistungsfähigeren Maschinen, die mit anspruchsvoller Sensorik und Motorik sowie vernetzt arbeiten, führt zu hybriden, soziotechnischen Konstellationen, in denen Menschen und Maschinen eng verwoben sind und auf komplexe Weise interagieren. Dies kann das Verhalten und die Handlungen von Menschen stark beeinflussen und ihre individuellen Freiheitsspielräume und Kontrollmöglichkeiten einschränken. Zudem können fortgeschrittene und mit flexiblen, selbstlernenden Algorithmen arbeitende maschinelle Systeme menschliches Tun zum Teil so gut imitieren, dass sie wie intentionales menschliches Handeln erscheinen, was weitere ethische Fragen aufwirft (vgl. Kapitel 4).

Auch mit Blick auf diese soziotechnische Komplexität bis hin zu Unterscheidungsschwierigkeiten erscheint es sinnvoll, an einem engen Handlungsbegriff, der an das zentrale Kriterium der Intentionalität gebunden ist, festzuhalten. Dieses Intentionalitätskriterium ist zudem entscheidend für die Möglichkeit der Zuschreibung von Verantwortung im Kontext von Mensch-Maschine-Interaktionen in zunehmend komplexer soziotechnischer Vernetzung. Im Hinblick auf digitale Techniken handeln Menschen nicht jederzeit in vollem Maße autonom, sondern verstehen oft weder die technischen Zusammenhänge noch haben sie immer umfassende Informationen oder hinreichende Wahlfreiheit, um bewusste Entscheidungen im Umgang mit der Technik treffen zu können. Solche

Konstellationen können vielfältig ethische Bedeutung entfalten, wenn es um Fragen der Verantwortung geht.

3.3.2 Verantwortung

Fragen der Verantwortung und Verantwortlichkeiten knüpfen an den Handlungsbegriff an. Die Rolle menschlicher Verantwortung für die kausale Verursachung von Handlungsfolgen kann in vielfacher Weise thematisiert, reflektiert, diskutiert und modifiziert werden. Bei der Verantwortung für die Resultate von Zuschreibungshandlungen in sozialen Kontexten geht es vor allem darum zu klären, welche Verantwortung von welchen Akteuren übernommen werden *soll*, um einer zunehmenden Verantwortungsdiffusion entgegenzuwirken.[101] Verantwortungszuschreibung und -verteilung erfolgen zu dem Zweck, Praxisfelder wie den Straßenverkehr, den Schulbetrieb oder den Umgang mit KI in verschiedenen Anwendungsbereichen so zu strukturieren und gegebenenfalls rechtlich zu regulieren, dass sich dadurch eine möglichst „gute Praxis" entfalten kann.

In den Zuschreibungen wird der Kreis der verantwortungsfähigen Individuen abgegrenzt, der Stellenwert der kausalen Verursachung geregelt und es werden Kriterien festgelegt, welche Voraussetzungen Menschen erfüllen müssen, um ihnen Verantwortung zuschreiben zu können. Somit stellt sich die Frage, wer für was direkt oder indirekt Verantwortung übernehmen kann oder soll, wenn Individuen, Gruppen und Institutionen aus und in verschiedenen Bereichen wie im Privatleben, in der Forschung, Wirtschaft und Politik sowohl miteinander als auch mit maschinellen Systemen und insbesondere KI-Systemen zusammenwirken.

Verantwortung kann als Konzept einer vielfachen Relation rekonstruiert werden. Im Kontext dieser Stellungnahme

101 Grunwald 2021.

erscheint eine fünfstellige Relation angemessen: Wer ist für was, gegenüber wem, vor wem und unter welcher Norm verantwortlich?[102] Allerdings sprechen wir auch von einer verantwortlichen Person; in diesem Falle ist der Begriff einstellig und der Zusammenhang zwischen dem einstelligen moralischen Begriff der verantwortlichen Person und den unterschiedlichen, meist mehrstelligen Kriterien der Verantwortungszuschreibung ist eine eigene philosophische und rechtstheoretische Thematik. Ohne die genuine individuelle moralische Verantwortung wären auch die weiter gehenden Ausdifferenzierungen gegenstandslos.[103]

Demnach kann man erstens ganz grundsätzlich beim Verantwortungssubjekt (*wer*) ansetzen, das Verantwortung übernehmen kann. Verantwortungssubjekte tragen Verantwortung, als Einzelperson oder Mitglieder eines Kollektivs, beispielsweise einer Institution. Davon zu unterscheiden ist zweitens das Verantwortungsobjekt (*was*), für das Verantwortung übernommen wird, zum Beispiel Handlungen sowie deren Gründe und Folgen.[104] Als drittes Relationselement werden die vom Handeln des Verantwortungssubjektes (direkt oder indirekt) Betroffenen benannt (*gegenüber wem*). Das vierte Relationselement bildet die Instanz, vor der die Verantwortung übernommen wird (*vor wem*). Das Gewissen als Inbegriff der praktischen Vernunft, andere Personen oder auch eine staatliche Rechtsgemeinschaft sind hier denkbar. Für eine normative Stellungnahme ist zudem ein fünftes Relationselement bedeutsam, nämlich die Regel oder das Prinzip, dem eine

102 Loh 2017.
103 Nida-Rümelin 2011.
104 Personen können auch für ein Unterlassen verantwortlich sein. Auch durch ein Unterlassen wird ein Ereignis in der Welt verursacht. Die Verantwortung kann aus einer Sonderbeziehung des Unterlassenden im Hinblick auf das dadurch beeinträchtigte Rechtsgut erwachsen („Garantenstellung") (Freund 1992). Ein „Jedermannsunterlassen", für das es auf keine solche Sonderbeziehung ankommt, lässt sich darüber hinaus auf allgemeine Solidaritätspflichten stützen (Frisch 2016).

verantwortliche Praxis gerecht werden sollte, zum Beispiel das Prinzip, andere nicht zu schädigen (*unter welcher Norm*).[105]

Vor allem das dritte Relationselement – die Betroffenen – ist nicht immer leicht einzugrenzen. Ungeachtet dessen sind deren Berücksichtigung und Einbezug gerade in Anbetracht des steigenden Einsatzes von KI-Systemen in vielen Gesellschafts- und Lebensbereichen von zentraler Bedeutung. An dieses Desiderat knüpfen sich auch Forderungen nach Transparenz und Nachvollziehbarkeit, welche eine Voraussetzung für die Beteiligung und Berücksichtigung von Betroffenen darstellen.

Über diese Konstellation hinaus ist in der Verantwortungsdiskussion zum wissenschaftlich-technischen Fortschritt die epistemologische Dimension zu bedenken. Denn Handlungsfolgen sind oft nur unter hohen und nicht eliminierbaren Unsicherheiten des Wissens antizipierbar.[106] Verantwortungszuschreibung muss daher die Dimension des Handelns unter Unsicherheit und damit die Risikothematik[107] berücksichtigen. Dies lässt auch noch einmal zwischen retrospektiver und prospektiver Verantwortung unterscheiden. Beim Blick auf vergangene Handlungen kann eine nachträgliche Verantwortungsübernahme unter Umständen angezeigt sein. Vor allem die prospektive Verantwortung unterliegt den gerade konstatierten Unsicherheiten.

Um Verantwortung im Zusammenspiel von Menschen und maschinellen Systemen näher zu betrachten, kann das fünfstellige Verantwortungskonzept technikspezifisch gerahmt werden. Ausgangspunkt ist hier, dass eine Verantwortungsübernahme (als *Verantwortungssubjekt*) nur Personen als verantwortlichen Wesen möglich ist, beispielsweise den

105 Damit bildet der Verstoß gegen eine Norm den Anknüpfungspunkt von Verantwortung. Normen schränken die Freiheit des Einzelnen ein und bedürfen daher der Legitimation. Im Recht kann insoweit auf den Grundsatz der Verhältnismäßigkeit (i.w.S.) zurückgegriffen werden (Dechsling 1989).
106 Grunwald 2014.
107 Nida-Rümelin/Schulenburg/Rath 2012.

Individuen, die Technik entwickeln und herstellen, die ihren Einsatz etwa in der Politik oder Unternehmen ermöglichen und fördern, oder denjenigen, die Technologien einsetzen. Das *Verantwortungsobjekt* ist dann je nach Rolle der Verantwortungssubjekte und ihrer Handlungen zu beschreiben: zum Beispiel Planen, Erfinden, Entwickeln oder Anwenden. Zu den *Betroffenen* können sowohl die von dem technischen Angebot direkt angesprochenen Personen(gruppen) gehören, zum Beispiel Angestellte in einem Krankenhaus, die mithilfe KI-gestützter Software Entscheidungen treffen, als auch weitere Personen wie zum Beispiel diejenigen, die auf Grundlage solcher Entscheidungen Diagnosen, Therapieempfehlungen oder sonstigen medizinischen Rat erhalten. Relevante *Instanzen* und relevante *Normen* sind hierbei verknüpft. Rechtliche Verantwortung besteht in letzter Instanz gegenüber der staatlich verfassten Gemeinschaft.

Moralische Verantwortung können nur natürliche Personen übernehmen, insofern sie über Handlungsfähigkeit verfügen, das heißt in der Lage sind, aktiv, zweckgerichtet und kontrolliert auf die Umwelt einzuwirken und dadurch Veränderungen zu verursachen. Träfe dies auch auf Maschinen zu, wären auch diese verantwortungsfähig. Dann müsste Maschinen der Personenstatus zugeschrieben werden, was jedoch weder aktuell noch angesichts der in absehbarer Zukunft erwartbaren qualitativen Entwicklungen maschineller Systeme angemessen wäre. Verantwortung kann daher nicht direkt von maschinellen Systemen übernommen werden, sondern nur von den Menschen, die in je unterschiedlichen Funktionen hinter diesen Systemen stehen, gegebenenfalls im Rahmen institutioneller Verantwortung. Auch wenn ein technisches System eingesetzt wird, um im Rahmen einer automatisierten Datenauswertung Schlussfolgerungen wie die Gewährung eines Kredites anzuwenden, ist es die Verantwortung des Menschen,

dieses System in einer ethisch vertretbaren Weise zu entwickeln und einzusetzen.[108]

Wer nun konkret als Verantwortungsträger fungiert, kann mit dem Konzept der *Multiakteursverantwortung* umrissen werden.[109] Kommt es bereits zu facettenreichen Verantwortungsgefügen, wenn man von nur drei prinzipiellen Ebenen möglicher Verantwortungszuschreibung ausgeht – Individuen, Organisationen und Staat –, so entsteht ein noch komplexeres Bild, wenn man Wechselwirkungen zwischen verschiedenen Akteuren aus diesen drei Ebenen berücksichtigt. Dies gilt erst recht, wenn diese Interaktionen zumindest teilweise von algorithmischen Systemen gestützt oder vermittelt werden, die mitunter für andere Beteiligte selbst autonom und kaum durchschaubar zu agieren scheinen.

Hier stellt sich die Frage, wie Verantwortung sinnvoll zwischen unterschiedlichen Beteiligten geteilt werden kann, zum Beispiel zwischen denjenigen, die maschinelle Systeme konzipieren und entwickeln, die ihre Nutzung beauftragen oder vorantreiben, die in Nutzungsprozesse oder ihre Überwachung direkt eingebunden sind, die Ergebnisse solcher Prozesse verwenden oder von ihnen direkt oder indirekt betroffen sind oder die ihre Auswirkungen auf unterschiedlichen gesellschaftlichen Ebenen beobachten und eventuell regulierend eingreifen können. In Anlehnung an das Konzept der Datensouveränität[110] ist die geeignete Gestaltung von Multiakteursverantwortung demnach zentral für eine angemessene informationelle Freiheitsgestaltung, die den Chancen und Risiken einer zunehmend digital vernetzten und algorithmisch gestützten Welt gerecht wird. Eine solche Freiheitsgestaltung kann nur dann verantwortlich sein, wenn sie sich gleichzeitig an den gesellschaftlichen Anforderungen von Solidarität und Gerechtigkeit orientiert.

---

108 Datenethikkommission der Bundesregierung 2019.
109 Deutscher Ethikrat 2017, 249 f.
110 Ebd., 252 f.

Auch für die Diskursführung über die zahlreichen Wechselwirkungen von Menschen und Maschinen und die gesellschaftlichen Auswirkungen einer zunehmenden Durchdringung der menschlichen Gesellschaft mit algorithmischen Systemen muss Verantwortung übernommen werden. Warnungen vor unkritischem Vertrauen in maschinelle Systeme, insbesondere im Falle Künstlicher Intelligenz, sollten einen Platz haben und sind Ausdruck wahrgenommener Verantwortung. Ebenso sind die Auswahl und Gewichtung bestimmter normativer Kriterien und Prinzipien im Diskurs zur Ethik von maschinellen Systemen, Algorithmen und KI selbst Gegenstand von Kontroversen.[111] Wer hat die Deutungshoheit, Werte und Normen, die im Umgang mit KI relevant sind, zu bestimmen? Wer prüft, welche Betroffenen vornehmlich in den Blick genommen werden oder wie die Lasten und Nutzen bestimmter Anwendungen verteilt sind? Es stellt sich also allgemein die Frage, wer für das fünfte Relationselement, die Bestimmung normativer präskriptiver Prinzipien, zuständig ist.

3.4 Anthropologische Aspekte des Mensch-Maschine-Verhältnisses

3.4.1 Philosophische Grundbestimmung des Menschseins

Handlung, Vernunft und Verantwortung stehen im Zentrum humanistischer[112] Philosophie. Menschen sind befähigt zur Handlungsurheberschaft und somit zur Autorschaft ihres Lebens. Sie sind frei und tragen daher Verantwortung für die Gestaltung ihres Handelns. Freiheit und Verantwortung sind

111 Jobin/Ienca/Vayena 2019; Rudschies/Schneider/Simon 2021.

112 Die unterschiedlichen Verwendungsweisen des Humanismusbegriffs in der Philosophie stimmen in einigen normativen Kernelementen überein, die zur Klärung anthropologischer Aspekte des Mensch-Maschine-Verhältnisses wichtig sind und nachfolgend näher entfaltet werden.

zwei einander wechselseitig bedingende Aspekte menschlicher Autorschaft. Autorschaft ist wiederum an Vernunftfähigkeit gebunden. Die strafrechtlichen Kriterien für Schuldfähigkeit konvergieren mit der lebensweltlichen Praxis moralischer Zuschreibungen. Personen sind jedenfalls in wichtigen sozialen Kontexten moralisch verantwortlich. Das heißt, man erwartet von ihnen, dass sie zurechnungsfähig handeln und urteilen.[113]

Diese Trias aus Vernunft, Freiheit und Verantwortung prägt heute sowohl die lebensweltliche Moral als auch die Rechtsordnung in hohem Maße. Im Mittelpunkt steht dabei das Phänomen der Affektion durch Gründe. Praktische Gründe sprechen *für* Handlungen, sie sind per se normativ, nicht erst über den Umweg eigener Wünsche. Ein Grund spricht dafür, das zu tun, was diesen Grund erfüllt, wenn nicht andere Gründe dem entgegenstehen.[114] Theoretische Gründe sprechen *für* Überzeugungen; auch diese sind normativ. In der Regel gibt es Gründe, das eine zu tun und das andere zu lassen, die gegeneinander abgewogen werden müssen. Der Konflikt von Gründen zwingt dann zur Abwägung und zur Systematisierung dieser Abwägung in Gestalt ethischer Theoriebildung.

Die menschliche Lebensform ist von reaktiven Einstellungen und moralischen Gefühlen geprägt, die von normativen Gründen begleitet sind. Wir vergeben einer Person, die uns Unrecht getan hat, wenn wir den Eindruck haben, sie habe das Unrechte ihres Tuns eingesehen und werde diese Praxis nicht fortsetzen. Wir sind dankbar, wenn wir meinen, dass eine Person etwas Gutes getan hat, ohne daraus Vorteile zu ziehen; wir nehmen etwas übel nur dann, wenn wir die betreffende Person

113 In der Technikanthropologie, die sich mit anthropologischen Aspekten des Technischen und der Technik befasst, vor allem mit Mensch-Technik- oder Mensch-Maschine-Konstellationen, wird eine Vielzahl auch anderer Perspektiven verfolgt (Heßler/Liggieri 2020). Hierzu gehören etwa Homo Faber und Homo Creator, trans- und posthumanistische Positionen sowie die Akteur-Netzwerk-Theorie. Die enge Verbindung ethischer Fragen zu den Konzepten von Freiheit und Verantwortung impliziert, diese in den Betrachtungen nicht eigens zu berücksichtigen, sondern die humanistische Perspektive in die Mitte zu stellen.

114 Scanlon 1998; 2014; Halbig 2007.

für voll zurechnungsfähig und in ihrem Handeln frei halten.[115] Die verobjektivierende Einstellung gegenüber anderen Menschen, die diese zum bloßen Gegenstand der Beeinflussung macht, sie gewissermaßen zu einem Teil der Umwelt degradiert, lässt sich nur für ganz spezifische Situationen – wenn überhaupt – durchhalten. Aber wenn diese verobjektivierende Einstellung ohne moralische Empfindungen zur allgemeinen Praxis würde, gäbe es die menschliche Lebensform nicht mehr. Diese ist gerade dadurch geprägt, dass wir ein Verhalten übel nehmen, wenn es uns inakzeptabel erscheint, dass wir zum Beispiel in der Lage sind zu verzeihen, wenn wir dafür Gründe haben, oder dass wir Dankbarkeit empfinden.

Freiheit kommt insofern ins Spiel, als wir *rationaliter* bestimmte moralische Gefühle und reaktive Einstellungen zurückstellen, wenn wir erfahren, dass die betreffende Person in ihrem Handeln nicht frei war, was immer die Ursachen dieser Unfreiheit sind, wie beispielsweise äußerer Zwang, psychische Erkrankung oder überwältigende Angst. Diese Praxis der Zuschreibung von Freiheit und Verantwortung ist essenziell für die Grundlegung moralischer Beurteilung wie auch für moralische Gefühle und reaktive Einstellungen. Daher ist es ausgeschlossen, diese aufzugeben, wie überzeugend auch immer wissenschaftliche Theorien, die prima facie dagegen sprechen, sein mögen.[116] Es macht unsere Zugehörigkeit zur menschlichen Lebensform aus, dass solche moralischen Beurteilungen, Gefühle und Einstellungen unsere soziale Praxis prägen. Die Normen von Moral und Recht sind ohne die Annahme menschlicher Verantwortlichkeit und damit von Freiheits- und Vernunftfähigkeit unbegründet. Sie würden in bloße Instrumente der Verhaltenssteuerung transformiert[117] und paradoxerweise wäre es gerade diese Transformation, die ihre

115 Vgl. den einflussreichen Ansatz von Peter Strawson (1974).
116 Siehe beispielsweise die Debatte rund um die Experimente von Benjamin Libet (2004).
117 Vgl. dazu Moritz Schlick (1930), der exemplarisch für diese antihumanistische Auffassung steht.

Wirksamkeit für die Verhaltenssteuerung zugleich gefährden würde.

Wenn menschliche Freiheit im Sinne des Anderskönnens bestritten wird, kann an der Verantwortlichkeit menschlicher Personen nicht festgehalten werden.[118] In § 20 StGB werden die praktischen und epistemischen Bedingungen von Schuldfähigkeit dargestellt, die ohne diese anthropologischen Voraussetzungen von Freiheit und Vernunft nicht aufrechtzuerhalten wären. Ohne die Fähigkeit, sich anders entscheiden zu können, gibt es keine Handlungsurheberschaft und keine Verantwortung.

Obwohl die humanistische Perspektive nicht nur die lebensweltliche Normativität, sondern auch die juridische Deliberation von den Menschenrechten bis zum Strafrecht imprägniert und das kulturelle Fundament demokratischer Ordnungen ausmacht, wird sie doch immer wieder infrage gestellt. Zwei jüngere Formen der Kritik, die sich teilweise überlagern, stützen sich einerseits auf neurowissenschaftliche Begriffe und Paradigmen sowie andererseits auf solche aus Debatten um das Thema KI. In den Neurowissenschaften wurden bestimmte empirische Studien, nach denen zum Beispiel das motorische Zentrum des Gehirns schon mit der Vorbereitung einer Bewegung beginnt, bevor man sich bewusst für die Ausführung der Bewegung entschieden hat[119], als Beleg dafür interpretiert, dass es Freiheit und damit menschliche Verantwortlichkeit nicht gebe. Stattdessen handele es sich dabei lediglich um – möglicherweise sozial nützliche – Illusionen. Tatsächlich lassen solche Befunde jedoch unterschiedliche Interpretationen zu und eignen sich nicht als Widerlegung menschlicher Freiheit und Verantwortlichkeit. Auch wenn alle mentalen Prozesse neurophysiologisch realisiert sind, sprechen die empirischen

118 Es gibt allerdings in der zeitgenössischen Philosophie ebenso die Auffassung, Verantwortlichkeit könnte auch ohne die Bedingung der Freiheit postuliert werden – besonders prominent bei Harry Frankfurt (Kane 2011, Kap. 5; siehe auch Frankfurt 1971).

119 Libet 2004.

Befunde aus den Neurowissenschaften nicht gegen das Phänomen normativer Gründe und ihrer Rolle für menschliche Handlungsmotivation, da wir es hier mit zwei Sprachebenen zu tun haben, die sich wechselseitig nicht in die Quere kommen können.[120]

Die zweite, von der KI-Debatte inspirierte Kritik der humanistischen Anthropologie changiert zwischen einer Überwindung des Menschen in Gestalt des Transhumanismus, der mit neuen Mensch-Maschine-Symbiosen die Reichweite menschlichen Wirkens in neue Dimensionen heben möchte, und einem Maschinenparadigma, das den menschlichen Geist auf das Modell eines algorithmischen Systems reduziert. Gerade Letzteres entfaltet besondere Relevanz im Kontext dieser Stellungnahme, da es großen Einfluss auf die Interpretation der Wechselwirkungen zwischen Mensch und Maschine und deren Rückwirkungen auf das menschliche Selbstverständnis hat.

3.4.2 Der Mensch als Maschine – die Maschine als Mensch?

Der Mensch als Maschine ist eine alte Metapher, deren Ursprünge bis in die frühe Neuzeit zurückreichen. Der mechanistische Materialismus des rationalistischen Zeitalters lässt die Welt als Uhrwerk erscheinen und den Menschen als Rädchen im Getriebe. Der große Uhrmacher ist dann der Schöpfer, der dafür gesorgt hat, dass nichts dem Zufall überlassen ist und ein Rädchen ins andere greift. Für menschliche Freiheit, Verantwortung und Vernunft ist in diesem Bild kein Platz.

Die vielleicht aktuell größte Herausforderung für das humanistische Menschenbild stellt das digital erneuerte Maschinenparadigma des Menschen dar. Das digitale Weltmodell, die

120 Strawson 1974; Nida-Rümelin 2005; Korsgaard 1996.

Welt als umfassender Algorithmus[121], scheint als zeitgenössische Variante des Maschinenparadigmas einer Deutung der Welt als Maschine eine attraktive Interpretation anzubieten. Diese beruht auf der Unterscheidung zwischen Software und Hardware. Es handelt sich dabei um zwei Beschreibungsebenen: die der Hardware, die lediglich auf physikalische und technische Begriffe zurückgreifen muss, und die der Software, die sich wiederum in eine syntaktische und eine semantische Ebene aufteilen lässt. Die syntaktische Beschreibung beruht auf der Zeichenverarbeitung, genauer dem Vokabular und den Regeln der Zeichenverarbeitung. Die Semantik unterlegt den Zeichen und den Regeln, nach denen sie verarbeitet werden, eine Bedeutung. Im Falle von Behauptungssätzen führt diese Unterlegung zu einer Wahrheitswertverteilung; die Sätze werden dann als wahr oder falsch markiert; im Falle einer arithmetischen Semantik folgt die Wahrheitswertverteilung mathematischen Regeln.

Die Beschreibung (und Erklärung) von Softwaresystemen als Hardware ist geschlossen: Jeder Vorgang (Ereignis, Prozess, Zustand) lässt sich als kausal determiniert durch den vorausgegangenen Zustand der Hardware eindeutig bestimmen. Als Modell auf den Menschen übertragen heißt dies, dass die physikalisch-physiologische „Hardware" wie ein algorithmisches System mit einer durch Genetik, Epigenetik und sensorische Stimuli eindeutig festgelegten zeitlichen Zustandsfolge von Eigenschaften funktioniert, die durch mentale Termini beschrieben wird und damit bedeutungsvolles Reden und Handeln ermöglicht. Das humanistische Menschenbild und damit die normativen Grundlagen von Moral und Recht erweisen sich dann als pure Illusion bzw. kollektive menschliche Selbsttäuschung.[122]

Schon in der ersten Digitalisierungswelle nach dem Zweiten Weltkrieg erwies sich interessanterweise nicht das eben

121 Nida-Rümelin/Weidenfeld 2018.
122 Singer 2004; Bennett et al. 2007.

geschilderte materialistische Maschinenparadigma, sondern eine animistische Variante als wirkungsmächtiger. Das animistische Paradigma geht gewissermaßen den umgekehrten Weg der Interpretation: Anstatt den menschlichen Geist und mentale Zustände als Epiphänomene materieller Prozesse in einer physikalisch geschlossenen Welt zu interpretieren und das Materielle mechanistisch zu beschreiben, wird nun im Kontext des Turing-Tests (vgl. Abschnitt 2.1) das algorithmische System mit mentalen Eigenschaften ausgestattet, sofern es in seinem äußeren (Ausgabe-)Verhalten demjenigen von Menschen hinreichend (das heißt verwechselbar) ähnelt.

Entsprechend war in der ersten Phase der Diskussion um KI ab den 1970er-Jahren die Frage, ob Computer denken können, leitend. Für die Fragerichtung ist die kontroverse Diskussion um die Interpretation des von Turing vorgeschlagenen Kriteriums paradigmatisch.[123] Wie zuvor bereits dargelegt, kann nach Turing die Frage, ob technische Artefakte „denken" können, dadurch entschieden werden, dass eine Person (für sie verdeckten) Menschen und Geräten beliebige Fragen stellt. Wenn in einer größeren Zahl von Durchgängen mit wechselnden Fragenden und wechselnden Menschen/Geräten die Antworten zu einem hinreichend großen Anteil (z. B. 50 Prozent) nicht eindeutig Mensch oder Gerät zugeordnet werden können, gibt es nach Turing keinen Grund, technischen Artefakten weniger Denkvermögen zuzuschreiben als Menschen.

Die in Diskursen rund um KI teilweise verbreitete Tendenz, im Anschluss an den Turing-Test eine äußerliche Ununterscheidbarkeit von menschlicher und maschineller Performanz pauschal mit der Annahme von Intelligenz und Denkvermögen solcher Maschinen gleichzusetzen, auf diese Weise die Differenz zwischen dem Simulierten und dem Simulierenden einzuebnen und die menschliche Vernunft damit tendenziell für maschinell ersetzbar zu halten, ist kein Zufall, sondern das Ergebnis bestimmter theoretischer Vorannahmen

123 Turing 1950; vgl. die kritische Darstellung bei Mainzer 1995, 113 f.

insbesondere *behavioristischer* und *funktionalistischer* Art.[124] Schon der klassische Behaviorismus[125] hatte sich zu Beginn des 20. Jahrhunderts in dem Bemühen, das menschliche Verhalten auf der Grundlage präzise beschreibbarer Reiz-Reaktion-Schemata zu erklären und die Psychologie damit in eine exakte Wissenschaft zu verwandeln, im Grunde einer Blackbox-Methode bedient, die das Innenleben derart beschriebener Organismen komplett ausblendet.

Der Text von Turing ist zweifellos vom logischen Behaviorismus inspiriert, der in den Nachkriegsjahren die zeitgenössischen Debatten insbesondere in der britischen Philosophie zunehmend prägte[126] und nach dem sich mentale Zustände ontologisch auf Verhaltensdispositionen reduzieren lassen, also auf die Neigung eines Organismus, sich unter bestimmten Bedingungen auf eine bestimmte Weise zu verhalten. Ein mentaler Zustand wie Schmerz ist demnach lediglich ein Verhaltensmuster, etwa die Veranlagung zu schreien oder zu weinen, wenn man sich verletzt hat. Auch Turings Text identifiziert den Sinn eines sprachlichen Ausdrucks nicht etwa mit der Intention der sprechenden Person, sondern mit den empirischen Verhaltensmustern, die mit einer Äußerung üblicherweise einhergehen. Die Paradoxa, die den logischen Behaviorismus unglaubwürdig machen, gelten auch für die Turing'sche Variante: Auch wenn wir die Bedeutung eines Satzes wie „Diese Person hat Schmerzen" lernen, indem wir darauf achten, welches Verhalten jeweils darauf hinweist, dass sie Schmerzen hat, so kann schon deshalb die Bedeutung von „Schmerzen haben" nicht lediglich ein Verhaltensmuster sein, weil „Superspartaner", die keine Schmerzen zeigen, dann auch keine Schmerzen haben könnten.[127] Obwohl sich der behavioristische Theorieansatz schon bald als zu eng erweisen sollte und in den

124 Bormann 2021, 51 ff.
125 Watson 1930.
126 Wittgenstein 1953; Ryle 1949. Einen guten Zugang zur damaligen Ordinary Language Philosophy vermittelt von Savigny 1974.
127 Putnam 1965.

folgenden Jahrzehnten verschiedene Transformationen erfuhr, fand er mit dem sich seit den 1950er-Jahren ausbreitenden Funktionalismus, nach dem mentale Zustände funktional vollständig erfasst werden, eine Nachfolgetheorie, die der aufkeimenden Kognitionspsychologie und Computerwissenschaft noch besser entsprach, weil Computer in dieser Lesart ein geeignetes Modell für mentale Prozesse sein können.[128]

Der Reiz funktionalistischer Ansätze besteht zunächst darin, dass sie sich gegenüber den ontologischen Implikationen des Leib-Seele-Problems neutral verhalten, das heißt, sie umgehen die Frage nach der Beziehung zwischen Körper und Geist sowie die Frage, wo und wie sich in diesen das denkende und fühlende Subjekt verorten lässt. Der Funktionalismus plädiert dafür, die Frage nach der Seinsart mentaler Zustände zugunsten der genauen Beschreibung ihrer Funktion aufzuheben. Durch die These der *multiplen Realisierung*[129], nach der bestimmte mentale Ereignisse, Eigenschaften oder Zustände durch ganz unterschiedliche physikalische Ereignisse, Eigenschaften oder Zustände realisiert werden können, schien es zudem möglich, auch Computern mentale Zustände zuzuschreiben, obwohl sie keine biologischen Strukturen besitzen.[130] Indem der Funktionalismus durch die ausdrückliche Anerkennung theorierelevanter innerer Zustände eines Systems nicht nur das Blackbox-Prinzip des Behaviorismus überwand, sondern mit der funktionalen Interpretation solcher Zustände auch die Integration biologischer und maschineller Entitäten in eine umfassende einheitliche Theorie zu ermöglichen schien, bahnte er insofern auch einer animistischen Deutung von KI-Systemen den Weg, als diese nun umso leichter als handlungsfähige Agenten mit einem mentalen Innenleben

128 Putnam 1975. Von dieser Position distanziert er sich zu späterer Zeit, siehe dazu Putnam 1992.
129 Für die unterschiedlichen Entwicklungsstufen dieser These vgl. Bickle 2020; Polger/Shapiro 2016.
130 Putnam 1967; Fodor 1974.

vorgestellt werden konnten, denen man zutraute, die menschliche Vernunft irgendwann einmal ersetzen zu können.

Obwohl der Funktionalismus aufgrund dieser Vorteile zunächst viel Zuspruch sowohl in der analytischen Philosophie des Geistes als auch in der KI-Forschung fand, wurden schon bald Einwände gegen dieses Theoriemodell vorgetragen. Eine erste gewichtige Kritik einer funktionalistischen Betrachtung des Mentalen, die den menschlichen Geist als Rechenmaschine[131] fasst, deren innere Zustände allein von ihrer Funktion im Sinne einer kausalen Verknüpfung von Eingabe und Ausgabe bestimmt werden, stammt von Thomas Nagel. Seines Erachtens lässt unsere gewöhnliche Auffassung mentaler Phänomene eine solche reduktionistische Sichtweise allein schon deswegen nicht zu, weil das Mentale neben seiner bloßen Funktion auch durch ein bestimmtes *phänomenales Bewusstsein*[132] geprägt sei, das in dieser Beschreibung verloren gehe. Nagel leugnet weder, „daß bewußte mentale Zustände und Ereignisse Verhalten verursachen, noch, daß man sie funktional charakterisieren könnte", sondern bestreitet lediglich, „daß derartiges eine vollständige Analyse ergibt".[133]

Ein Wesen kann nur *als* dieses Wesen mentale Zustände haben. Daher kann von den *Empfindungsqualitäten* des phänomenalen Bewusstseins nicht abgesehen werden, die allein aufgrund *äußeren* Verhaltens nicht zugänglich sind. Es fühlt sich für uns immer auf eine ganz bestimmte Art und Weise an, ein erlebendes, denkendes und handelndes Subjekt zu sein. Diese besonderen Qualitäten sind kein flüchtiges Beiwerk mentaler Zustände, sie gehören vielmehr insofern konstitutiv zu allen unseren Erfahrungen, als wir die uns umgebende Welt prinzipiell gar nicht anders erleben können als aus der Perspektive eines solchen phänomenalen Bewusstseins.

---

131 Oder – um eine Kritik von Ned Block (1978) aufzugreifen – wie einen Getränkeautomaten („Coke machine").

132 Zur Debatte um die Bedeutung des phänomenalen Bewusstseins vgl. auch Heckmann/Walter 2006; Metzinger 2009.

133 Nagel 1981, 262.

Dieses phänomenale Bewusstsein setzt dem Vermögen, die Qualität des Erlebens oder die mentalen Zustände anderer Lebewesen zu beurteilen, gewisse Grenzen. Dies illustriert Nagel am Beispiel des Empfindens einer Fledermaus, deren Orientierung aufgrund spezifischer sensorischer Besonderheiten (wie Radar oder Echolotortung) ganz anders strukturiert ist als beim Menschen. Als Menschen können wir zwar versuchen, uns vorzustellen, wie es ist, sich auf gänzlich andere Weise im Raum zu orientieren, wir bleiben dabei aber immer unserer eigenen, spezifisch menschlichen Weise des Erlebens verhaftet, ohne jemals Zugang zu den besonderen Qualitäten der mentalen Zustände einer Fledermaus zu erhalten. Unsere subjektive Perspektive bleibt unüberwindlich.[134] Vor diesem Hintergrund dieses auch als Qualia-Argument[135] bezeichneten Gedankengangs basiert die funktionalistisch inspirierte Mensch-Computer-Analogie auf einer fragwürdigen Reduktion, die nur gelingen kann, „wenn die artspezifische Betrachtungsweise von dem, was reduziert werden soll, ausgeklammert wird".[136]

Gegen eine funktionalistische Interpretation Künstlicher Intelligenz hat John Searle ein Argument entwickelt, das als das meistdiskutierte in der zeitgenössischen Philosophie gilt: das Chinese-Room-Argument.[137] Es verweist auf ein Gedankenexperiment, in dem eine Person in einem Zimmer sitzt, in das durch einen Schlitz jeweils Fragen in chinesischer Schrift gereicht werden. Die Person reicht Antworten auf diese Fragen ebenfalls durch den Schlitz heraus. Wenn die Antworten hinreichend plausibel erscheinen, mag man vermuten, dass die Person im chinesischen Zimmer des Chinesischen mächtig ist. Nun stellt sich aber heraus, dass jede der eingereichten Fragen eine Ziffer trägt und die Person über vorgefertigte Antworten und eine Tabelle mit Zuordnungen verfügt, sodass sie lediglich

134 Ebd.
135 Von lat. *qualis* = wie beschaffen.
136 Nagel 1981, 269.
137 Searle 1980. Eine frühere Zurückweisung des Funktionalismus stammt von Block 1978.

eine Antwort heraussuchen muss, die eine Ziffer trägt, die der Ziffer der Fragestellung zugeordnet ist. Die Person beherrscht die chinesische Sprache nicht, auch nicht das Zimmer als Ganzes mit einer Eingabe- und Ausgabefunktion. Aber zweifellos muss es irgendjemanden geben, der des Chinesischen mächtig ist und daher in der Lage war, den Fragen mithilfe der Ziffern Antworten so zuzuordnen, sodass der Eindruck entsteht, dass die Person im chinesischen Zimmer Chinesisch versteht. Die Analogien zu Softwaresystemen liegen auf der Hand. Es handelt sich um Zuordnungsregeln (genauer um Algorithmen), die lediglich für die Programmierung und den Gebrauch der digitalen Maschine Bedeutung haben, aber nicht das Softwaresystem selbst zu einer semantischen Maschine machen. Dieses verfügt nicht über Bedeutungen, es versteht nichts, es entscheidet nichts. Softwaresysteme verfügen nicht über eine Semantik, es handelt sich nicht um semantische Maschinen.

3.4.3 Verleiblichte Vernunft

Die Zurückweisung funktionalistischer Maschinenparadigmen lenkt den Blick auf die bereits im Qualia-Argument angedeutete besondere Qualität menschlicher Vernunft und deren Bedeutung für das menschliche Selbstverständnis. Menschliche Vernunft ist leibliche Vernunft. Diese Einsicht wendet sich gegen den in der abendländischen Tradition lange herrschenden Gedanken eines Dualismus zwischen Natur und Vernunft, Körper und Geist, der den Menschen als Naturwesen auf der einen und als Vernunftwesen auf der anderen Seite begreift. Solch dualistische Vorstellungen wurden sowohl durch Erkenntnisse in der Evolutionsbiologie als auch in der Hirnforschung und in den Kognitionswissenschaften infrage gestellt, die stattdessen auf die Relevanz des Leiblichen für die Bestimmung menschlicher Intelligenz und auf die Bedeutung unbewusster Prozesse für die Entwicklung höherer geistiger Leistungen verweisen.

Mit Maurice Merleau-Pontys Unterscheidung von zweierlei Arten des Körpers kann dies veranschaulicht werden.[138] Er unterscheidet den lebendigen handelnden *Leib* vom rein physikalischen *Körper*. Die Fähigkeit, soziale Bindungen einzugehen, sich in andere hineinzuversetzen, wird ermöglicht dadurch, dass der Mensch Leib ist und nicht nur einen Körper hat. Mit dem Leib ist der empfindende Organismus gemeint, mit seinem Vermögen, zu fühlen und sich zu bewegen. Wir sind an diesen Leib gebunden, in allem, was wir denken und tun. Er ist daher Ausgangspunkt und Bestandteil jeder Wahrnehmung und Empfindung. Als solcher ist er Voraussetzung für unser In-der-Welt-Sein und zugleich dafür, eine Welt zu haben und eine Beziehung zu anderen herzustellen.

Kognitive Fähigkeiten sind in ihrem Entstehungs- und Vollzugsprozess also an Sinnlichkeit und Leiblichkeit gebunden. Dies hat Konsequenzen für unser Verständnis vom Gehirn als Erkenntnisorgan und von der Vernunft als Erkenntnisvermögen – und damit auch für das Verständnis unseres Zugangs zur Realität. Wesentlich ist dabei, dass das in der Verkörperung des Gehirns eingeschlossene Naturverhältnis einer leiblichen Vernunft des Menschen seine Sozialität impliziert und seine Kulturalität bestimmt. Im menschlichen Leib sind Sozialität und Kulturalität von Anfang an angelegt, vor aller Entwicklung eines reflexiven und sprachlich vermittelten Bewusstseins.[139] Denn leibliche Vernunft vollzieht nicht nur einen kognitiven Informationsaustausch, sondern mit ihr spielen auch Kommunikation und Kooperation eine Rolle.[140]

---

138 Merleau-Ponty 1966.

139 Fuchs 2013, 11 ff. Vier Erscheinungsformen des Leibes können diese Anlage der Vermittlung zwischen der Natur- und der Kulturseite des Menschen plausibel machen: (i) ein mit der Umwelt vertrauter Leib, der sich vor allem im Umgang mit kulturellen Gegenständen entwickelt, (ii) ein „passiv-affizierbarer" Leib, der affektiv mit anderen verbunden ist, (iii) ein „mimetisch-resonanter" Leib, der durch Nachahmung in grundlegende Kommunikation mit anderen eingebunden ist, bis er so als (iv) kooperativ kultivierter Leib zum Körper für andere wird, indem er Haltungen und Rollen übernimmt, die ihm somit zur „zweiten Natur" werden (ebd., 26 f.).

140 Schmitz 1990.

Beides sind Faktoren, die von Kindheit an entscheidend sind für jene bewussten Prozesse, in denen sich die Kulturfähigkeit bildet, die dem Menschen als sozusagen „zweite Natur“[141] zuwächst. Für das menschliche Gehirn bedeutet das, dass es mit all seinen sich entfaltenden Fähigkeiten von Anfang an in biologisch-organische wie in sozial-kulturelle Entwicklungsprozesse eingebunden ist.

Ein solches Verständnis des Menschen geht nicht von einer blutleeren „reinen“ Vernunft aus, sondern begreift auch die Vernunft als immer schon leiblich eingebunden und sozial wirksam. Damit ist die Frage nach dem Praktischwerden der Vernunft, das heißt nach den normativen Orientierungen und nach der Motivation moralischen Handelns, keine zweite Frage, sondern sie begleitet alles Denken, das als solches Lebensgestaltung nicht nur ermöglicht, sondern immer bereits vollzieht.

Bereits die praktische Einsicht, dass bestimmte normative Gründe für eine Handlung sprechen, wird damit *handlungswirksam*. Damit wäre der entscheidende Aspekt einer umfassenden Theorie praktischer Vernunft, nämlich die Frage, wie sich moralische Überzeugungen in Handlungen überführen lassen, berührt. Bezüglich dieser Frage plädiert der Philosoph John McDowell dafür, in diesem Kontext nicht entweder die Vernunft oder aber die subjektiven Einstellungen und Strebungen zum alles beherrschenden Faktor der Handlungsverursachung zu stilisieren, sondern zu akzeptieren, dass in der menschlichen moralischen Erfahrung diese beiden (das heißt die kognitive und die appetitive Dimension) immer schon unauflöslich miteinander verschränkt seien.[142]

Entscheidend ist dabei, dass leibliche Erfahrungsstrukturen einhergehen mit der Fähigkeit, sich in andere hineinversetzen und sich mitteilen zu können, das heißt mit einer Prosozialität,

141 McDowell 2001, 109; 2002a.
142 McDowell 2002b, 177.

auf deren Grundlage sich die Fähigkeit zu geteilter Intention entwickeln und Empathie und Motivation initiieren kann.

Insofern nun das Gehirn kein isolierter Gegenstand ist, sondern eingelassen ist in Erfahrungen gemeinsamer Praxis, in der sich körperlich-biologisches und kulturell-soziologisches Erleben verschränken, entwickeln sich in solch sozialer Praxis ein Bedeutungswissen und ein Wissen um die Perspektivität von Erkenntnis, die reflexiv auf die Relationalität von Wissen und Erkenntnis verweist. Denn diese Perspektivität des Wissens erschließt sich insbesondere durch dessen In-Relation-Stehen zum eigenen Leib, durch die sogenannte „Eigenleiberfahrung". So wird mit dem Rekurs auf die verleiblichte Vernunft deutlich, dass zur menschlichen Intelligenz unabdingbar Reflexivität hinzugehört. Diese setzt menschliche Erkenntnis instand, zu unterschiedlichen Perspektiven Stellung nehmen und urteilen zu können.

Grenzen der Formalisierbarkeit und Simulierbarkeit menschlicher Vernunft

Mit der Reflexivität des Bewusstseins ist das Verstehen- und Vermittelnkönnen angesprochen, mit anderen Worten die hermeneutische Dimension, die sich auch in der Unterscheidung und Anwendung verschiedener Wissensformen darstellt und die ein besonderes Charakteristikum menschlicher Intelligenz bildet. Diese hermeneutische Dimension von Wissen ist aber nur begrenzt formalisierbar oder simulierbar und bezieht sich auf den Sinn und die Bedeutung menschlichen Erkennens und Handelns. Die Aneignung menschlicher Erfahrung ist immer mit Deutungsprozessen verbunden und setzt immer ein Beteiligtsein, ein Engagement voraus.[143] Die Art und Weise, *wie* wir wissen (*knowing how*), ist eine eigene Kompetenz, die sich nicht durch bloßes Sachwissen (*knowing that*) abbilden lässt.

143 Meyer-Drawe 2001.

Auch hier spielt der Leib eine wichtige Rolle, denn er ermöglicht ein Handeln, das allein mittels bewusster Planung und Berechnung so nicht möglich wäre. In der leiblichen Verfasstheit gründet daher auch die Nichtsimulierbarkeit des Denkens. Mit der leiblichen Verankerung des Bewusstseins, die eine Komplexität von Hintergrunderfahrungen mit sich führt, die Voraussetzung für alle bewussten Prozesse der Planung und Entscheidung sowie deren Begründung bilden, stößt die Entwicklung von KI an ihre Grenzen.

Der Sachverhalt solch leiblich verfassten Hintergrundwissens bedeutet mithin eine Grenze des rationalistischen Versuchs, Wissen vollständig in formalisierte Regeln zu überführen und künstlich nachzubilden. Es zeigt sich in der leiblichen Verschränkung von kognitiven und emotional appetitiven Momenten im Vernunftvollzug dann vielmehr die Relevanz des Nichtformalisierbaren. Es geht im Begreifen nicht mehr nur darum, das *Was* – die Fakten – zu begreifen, sondern darum, *wie* wir verstehen. Und dies wird entscheidend durch unsere leiblichen Vollzüge und Fähigkeiten bestimmt, die vorbegrifflich und unausgesprochenen unser Verhalten mitbestimmen.

Menschliches Wissen ist insofern eingebettet in einen Horizont des Nichtwissens. Im Raum individueller und soziokultureller Erfahrung wird deutlich, dass nicht nur die Klarheit logischen Schließens, sondern auch die Vagheit und Offenheit menschliches Denken auszeichnet. Gerade Vagheit und Unbestimmtheit des Wissens sind Voraussetzung für Kreativität und Intuition, die ein Handeln unter der Bedingung von Ungewissheit ermöglichen, mit der Menschen situativ auf konkrete Herausforderungen reagieren und Verantwortung übernehmen können. Für menschliche Intelligenz ist kennzeichnend, dass sie sich auf plötzliche Situationen einstellen kann, um in Erlebnisgegenwart Entscheidungen für oder gegen Zukunftsszenarien zu treffen.

Wesentlich ist daher für menschliche Intelligenz und deren Verantwortungsfähigkeit auch das Erleben von und der Umgang mit Zeit. Entscheidungen werden in Gegenwart erlebt

und in solchem Gegenwartserleben bewusst gehalten. Auch dieses Gegenwartserleben ist in der Leiblichkeit der Vernunft verankert. Veranschaulicht werden kann dies an der Bedeutung des menschlichen Gedächtnisses. Das menschliche Gedächtnis funktioniert nicht wie ein Speicher, der einen gedanklichen Bestand bildet und abrufbar wäre. Vielmehr ist es durch Prozesse des Erinnerns und Vergessens ausgezeichnet. Was jeweils im Moment erinnert – oder vergessen – wird, hängt von der je konkreten leiblichen Verfasstheit und den sozialen Bezügen, in denen der Mensch steht, ab. Erinnern ist damit nicht gleichbedeutend mit dem Abrufen einer Information. Es ist vielmehr ein hermeneutischer Akt, mit dem sich ein Erfahrungsraum[144] vergegenwärtigt.

3.5 Fazit

Aus den vorherigen Überlegungen lassen sich einige entscheidende Aspekte kognitiver Leistungen und Operationen von Menschen und Maschinen zusammenfassen. Das Kognitive ist im Falle menschlicher Intelligenz unauflöslich mit den kognitiven und emotiven, ästhetischen und ethischen, technischen und gestalterischen, sozialen und individuellen sowie zeitlichen Dimensionen der menschlichen Lebenswelt verbunden. Menschliche Intelligenz zeigt sich nicht nur in einem kognitiv kohärenten Urteil, sondern auch in einer kohärenten Praxis. Diese ist gründegeleitet und Ausdruck von akzeptierten Werten und Normen, die nicht beliebig zur Disposition stehen. Der Mensch ist durch die Fähigkeit, Gründe zu geben und zu nehmen und sich im Urteil und im Handeln an diesen zu orientieren, als Spezies charakterisiert. Veränderungen im normativen Gefüge der eigenen Praxis bedürfen der Begründung und bedrohen im Grenzfall die persönliche Integrität und Identität. Ein hinreichend entwickeltes lebensweltliches

144 Ebd.

Orientierungswissen ist Voraussetzung für eine intelligente Praxis. Damit sich dieses aufbaut, muss die betreffende Person die Fähigkeit haben, Wichtiges von Unwichtigem zu unterscheiden und normative Grenzen zu akzeptieren.

Menschliche Intelligenz ergibt sich zudem nicht allein aus dem Orientierungsbedarf des Individuums in einer natürlichen Welt, sondern ist das Ergebnis sozialer Interaktion. Von Geburt an hängt das Wohlbefinden menschlicher Wesen und hängen deren Entwicklungschancen vom Austausch mit anderen Menschen ab. Der intelligente Umgang mit den Herausforderungen der Welt ist nicht das Ergebnis eines fortgesetzten Puzzlespiels, sondern im Wesentlichen Folge der Einbettung der eigenen individuellen Praxis in den größeren sozialen und kulturellen Zusammenhang. Mit dem Erwerb der Sprache können Kinder auf Gründe reagieren, sich von Gründen affizieren lassen und diese selbst auf ihr eigenes Handeln applizieren. Das kulturelle Wissen wird über diese Praxis von einer Generation auf die nächste übertragen, immer wieder veränderten Bedingungen angepasst und bettet das einzelne Individuum in die menschliche Lebensform ein. Im Ergebnis verwebt sich dann die Einheit der Person mit der Einheit des Wissens und der Einheit der menschlichen Lebensform. Die einzelne Person zerfällt nicht in Funktionalitäten, sondern wird zusammengehalten durch Gründe, die ihre theoretische und praktische Lebensorientierung bestimmen. Das Individuum wird zur Person und zum Handelnden. Die Identität der Person äußert sich in einer kohärenten Praxis, die von stabilen Gründen geleitet ist. Diese integriert unterschiedliche Aspekte menschlicher Existenz – kognitive, emotive, soziale, ethische, ästhetische, technische und gestalterische.

Es ist fraglich, ob eine derart gründegeleitete, multidimensional bestimmte und soziokulturell eingebettete kohärente Praxis selbst für komplexe maschinelle Systeme jemals plausibel sein könnte. Softwaresysteme leisten Beachtliches. In vielen Bereichen sind sie menschlichen Fähigkeiten bei Weitem überlegen. Aber sie verfügen nicht über ein Analogon zu

menschlicher Intelligenz. Es wird der Softwareentwicklung der Zukunft vermutlich in wachsendem Umfang gelingen, menschliche Fähigkeiten zu simulieren und in vielen Fällen zu übertreffen. Das sollte uns aber nicht dazu verführen, ihnen personale Eigenschaften zuzuschreiben, die für genuine menschliche Existenz essenziell sind.

Trotz dieser kategorialen Unterschiede von Mensch und Maschine beeinflussen Mensch und Maschine einander fortwährend. Menschen entwickeln zu bestimmten Zwecken Technologien, die auf die Handlungsmöglichkeiten von Menschen zurückwirken, indem sie jene verändern, erweitern oder vermindern. Diese Mensch-Technik-Relationen und ihre ethische Relevanz genauer zu bestimmen, ist Gegenstand des folgenden Kapitels.

4 MENSCH-TECHNIK-RELATIONEN

4.1 Einleitung

Das Verhältnis zwischen menschlichem Handeln, der Verfügbarkeit von Technik und technischen Prozessen ist für die Ethik hoch relevant, denn in diesem Verhältnis können sich Einfluss- und Randbedingungen für Autonomie und Freiheit des Menschen und damit die Möglichkeit der Zuschreibung von Verantwortung auf durchaus komplexe Weise ändern. Dies gilt vor allem und auf spezifische Weise bei KI-Systemen. Es ist daher im Rahmen einer ethischen Analyse und Beurteilung relevant, das Zusammenspiel von Mensch und Technik bzw. von menschlichem Handeln und technischen Prozessen differenziert zu erfassen. Menschen entwickeln und gestalten Technik und nutzen technische Produkte und Systeme oder darauf aufbauende Dienstleistungen als Mittel zum Zweck. Gleichzeitig wirken diese häufig zurück und beeinflussen menschliche Handlungsmöglichkeiten, von der Eröffnung neuer Optionen und der Vergrößerung von Freiheitsgraden bis hin zur Anpassungserzwingung. In diesem Kapitel soll es darum gehen, in welcher Art und Weise verschiedene Mensch-Technik-Relationen die Handlungsmöglichkeiten des Menschen erweitern oder vermindern können, bis hin zur Ersetzung menschlicher Handlungen durch maschinelle Vollzüge. Damit verbunden ist die Frage, wie sich die Spielräume für die Entfaltung menschlicher Autorschaft und die Übernahme von Verantwortung jeweils verändern.

Zum einen geht es darum, dass Tätigkeiten, die vormals (allein) von Menschen durchgeführt wurden, graduell an technische Systeme delegiert werden. Dies reicht vom Delegieren einfacher Tätigkeiten über das Automatisieren komplexer Tätigkeiten oder ganzer Funktionen bis hin zur vollständigen Ersetzung des Menschen durch Technik. Der Begriff „*ersetzen*“ beschreibt hier also den Endpunkt einer vollständigen

Delegation. Zum anderen geht es um Rückwirkungen dieses mehr oder minder umfassenden Delegierens auf menschliche Akteure, das heißt um Fragen, inwiefern jenes Delegieren Handlungsfähigkeit, Möglichkeiten, Fertigkeiten und Kompetenzen von Menschen *erweitert* oder *vermindert*.

Die drei Begriffe des Erweiterns, Verminderns und Ersetzens dienen in diesem Kapitel als analytische Matrix. Sie werden in den folgenden Kapiteln auf ausgewählte Sektoren bezogen, um ein differenziertes Bild der Veränderungen durch KI und ihrer ethisch relevanten Aspekte zu gewinnen.

4.2 Technikdeterminismus versus Sozialkonstruktivismus

Mensch-Technik-Relationen sind Gegenstand vieler Disziplinen. Verortet zwischen Informatik und Psychologie, beschäftigt sich insbesondere das Feld der Human-Computer Interaction bzw. Computer-Human Interaction mit dem Verständnis und der Gestaltung von Mensch-Maschine-Schnittstellen. In den Geistes- und Sozialwissenschaften haben insbesondere die Wissenschafts- und Technikforschung, die Science and Technology Studies, die Techniksoziologie und die Technikphilosophie Konzepte und Theorien zur begrifflichen Analyse von Mensch-Technik-Relationen bereitgestellt. Das Verhältnis von Menschen und Gesellschaft zur Technik wurde vielfach entlang der Deutungslinie zwischen sozialem Konstruktivismus und technologischem Determinismus beschrieben.[145] Dahinter steht die Frage nach dem letztlich treibenden Faktor: Folgen Technikgestaltung im Einzelnen und damit auch der technische Fortschritt als Prozess eher menschlich gesetzten Zwecken oder eher einer Eigendynamik, der sich Mensch und Gesellschaft letztlich unterordnen und anpassen müssen. Auch wenn es keine einheitliche Verwendungsweise dieser beiden

145 Grunwald 2007.

Deutungen gibt und auch wenn viele Ansätze keine der Extrempositionen vertreten, sondern sich an unterschiedlichen Stellen zwischen Sozial- und Technikdeterminismus verorten, ist eine kurze Erläuterung illustrativ und inhaltlich für diese Stellungnahme wichtig.

In der bereits seit den 1920er-Jahren vertretenen technikdeterministischen Sichtweise wird eine Eigenlogik im technischen Wandel vermutet, die Mensch und Gesellschaft zur Anpassung nötigt. Während sich einzelne Techniken auf menschliche Zwecke zurückführen lassen, folge die gesamte Technologieentwicklung einer inneren und damit nicht oder kaum beeinflussbaren Dynamik. Als Treiber hinter dieser vermuteten Eigendynamik wird immer wieder auf ökonomische Verhältnisse und insbesondere den wirtschaftlichen Wettbewerb zwischen Unternehmen, aber auch den Wettbewerb zwischen Staaten und Weltregionen um vordere Plätze in der technologischen Forschung und Entwicklung hingewiesen. Der auf diese Weise zustande kommende, sozusagen blinde technische Wandel wirke sodann mit seinen Produkten auf die Gesellschaft ein und führe zu Anpassungsnotwendigkeiten, die von konkreter Akzeptanz einzelner Techniken bis hin zur Adaptation an letztlich technologisches Denken reichen.[146]

In der sozialkonstruktivistischen Sichtweise dagegen treten Technologien nicht eigendynamisch oder zwangsläufig auf den Plan, sondern sind das Ergebnis komplexer und sozial situierter Entwicklungs- und Gestaltungspraktiken bzw. von Ko-Konstruktionsprozessen unter Mitwirkung zahlreicher Akteure. Die Technikgeneseforschung[147] hat herausgearbeitet, nach welchen Mechanismen Technik durch Entscheidungsprozesse aus ersten Ideen über Entwicklungsprogramme, Experimente und Prototypen bis zum letztendlichen Ergebnis entsteht. Beispielsweise wurde die Rolle von gesellschaftlichen Leitbildern

146 Rapp 1978; Ropohl 1982; Grunwald 2019.
147 Bijker/Hughes/Pinch 1987; Weyer et al. 1997; Weingart 1989.

in diesen Prozessen untersucht.[148] Sozialkonstruktivistisch gesehen werden Algorithmen, Roboter, digitale Dienstleistungen oder Geschäftsmodelle für digitale Plattformen von Menschen in möglicherweise langwierigen Entscheidungsprozessen und Handlungssträngen erfunden, entworfen, hergestellt und eingesetzt sowie weiterentwickelt und an neue Umgebungen angepasst. Die „Macher" der Digitalisierung arbeiten in der Regel in Unternehmen, Forschungsinstitutionen oder Geheimdiensten mit bestimmten Agenden, Interessen und Geschäftsmodellen. Wenn *andere* Personen und Institutionen mit *anderen* Werten, Perspektiven und Interessen mitgestalten könnten, würden beispielsweise KI-Systeme mit anderen Eigenschaften entstehen, als wenn man diese den einschlägigen Konzernen mit deren Interessen und Geschäftsmodellen überlässt. Diese Sicht eröffnete Gestaltungsmöglichkeiten und motivierte partizipative Ansätze in der Technikgestaltung wie beispielsweise das Constructive Technology Assessment.[149]

Die grobe Skizze dieser beiden Positionen macht die jeweiligen blinden Flecken deutlich. Weder darf die Bedeutung gesellschaftlicher Hintergründe oder spezifischer Entscheidungen bei der Entwicklung von Technik ignoriert werden, so etwa in Unternehmen oder der öffentlich geförderten Forschung, noch kann abgestritten werden, dass Technologien auf gesellschaftliche Realitäten und menschliche Handlungsmöglichkeiten zurückwirken. Daher erscheint es ratsam, Technikdeterminismus und Sozialkonstruktivismus als Pole eines empirisch vielfältigen und differenzierten Spektrums zu begreifen, die den Blick auf unterschiedliche Aspekte von Technikentwicklung und Mensch-Technik-Relationen werfen.[150] Entsprechend sind weder die technikdeterministische noch die konstruktivistische Betrachtung der Mensch-Technik-Relation falsch, beide sind jedoch unterkomplex. Sie werden

148 Dierkes/Hoffmann/Marz 1992.
149 Rip/Misa/Schot 1995.
150 Dolata/Werle 2007.

empirisch unzutreffend, wenn sie in ihrer jeweiligen Perspektive verabsolutiert werden. Die Mensch-Technik-Relation unterliegt vielmehr von Grund auf einem Ko-Konstruktionsverhältnis und kann als Ko-Evolution beschrieben werden.[151] Soziale Kontexte und normative Kriterien auf der einen und Technologien auf der anderen Seite entwickeln sich weiter in gegenseitiger Wechselwirkung. Die Verfügbarkeit von Technik beeinflusst Handlungsmöglichkeiten und deren Realisierung, aber auch die Bedingungen und Möglichkeiten menschlicher Weltwahrnehmung, wodurch sich Lebensstile und Einstellungen verändern können. Umgekehrt entstehen, wie dies die Technikgeneseforschung in vielen empirischen Studien belegt hat, neue Techniken vor dem Hintergrund von sozialen Befindlichkeiten, normativen Kriterien und Zukunftsentwürfen.

Gerade im Kontext der KI sind die Arbeiten der Anthropologin Lucy Suchman von großer Bedeutung. Ihr Buch „Human-Machine Reconfigurations"[152] liefert eine Reflexion und Kontextualisierung ihrer Studien der KI-Forschung in den 1980er-Jahren, welche 1987 unter dem Titel „Plans and Situated Actions"[153] veröffentlicht wurden. Suchman kritisiert das Planungsmodell von Interaktion, das einem Großteil der damaligen Forschung zugrunde liegt, und schlägt einen Perspektivenwechsel in der Betrachtung der Mensch-Maschine-Relation vor, der Einsichten aus den Sozialwissenschaften Rechnung trägt. Danach ist menschliches Handeln auf vielfältige Weise sozial situiert und beeinflusst, ohne vollständig determiniert zu sein. Sie argumentiert, dass Menschen sinnvoll handeln, indem sie auf der Grundlage ihrer sozialen und ökologischen Ressourcen häufig weniger planen als improvisieren. Sie kritisiert also die theoretisch-konzeptionellen Grundlagen des Designs interaktiver technischer Systeme als aus anthropologischer Sicht unangemessen, weil menschliches Handeln

151 Rip 2007.
152 Suchman 2007.
153 Suchman 1987.

ständig aus dynamischen Interaktionen mit der materiellen, insbesondere technischen und der sozialen Welt konstruiert und rekonstruiert werde.

Technikphilosophie und -soziologie haben die zunehmende Komplexität der Mensch-Maschine-Relation in unterschiedlichen Theorien gedeutet und zugespitzt. In der Technikphilosophie wird Technik häufig nicht mehr als Ensemble technischer Objekte verstanden, sondern als Medium, mit dem sich menschliches Handeln und Verhalten vollzieht. Während die einzelnen Elemente dieses Mediums instrumentellem Zweck-Mittel-Denken entstammen, stelle ihre Gesamtheit eine „zweite Natur" dar, die Rand- und Erfolgsbedingungen für weiteres menschliches Leben setzt und auch die Weltsicht und das Problemlösen beeinflusst.[154] Als die bereits technologisch orientierten Menschen, zu denen sie im Rahmen vieler Technisierungsprozesse geworden sind, werden sie zum Beispiel dazu neigen, Herausforderungen von Kommunikation oder Sicherheit als Probleme anzusehen, die primär technologisch zu lösen sind. Somit ist neue Technologie oft bereits das Ergebnis einer technologischen Art und Weise, wie Menschen die Welt sehen und sich zu ihr in Beziehung setzen.

Techniksoziologisch sind hier vor allem Ansätze der Ko-Evolution von Technik und Gesellschaft zu nennen.[155] In ihnen sind die sozialkonstruktivistischen Motive der Gestaltung aufgenommen, jedoch wird ihnen die vielfältige Rückwirkung einmal entwickelter und verfügbarer Technik auf Mensch und Gesellschaft zur Seite gestellt, zu denen beispielsweise die großen Infrastruktursysteme – wie jene der Mobilität – geeignetes Illustrationsmaterial liefern. Zunächst gestaltet nach Zweck-Mittel-Erwägungen unter Berücksichtigung vielfältiger gesellschaftlicher Belange beispielsweise aus Wirtschaft, Bürgerschaft und Umweltschutz, werden sie nach Fertigstellung zu Randbedingungen menschlicher Entscheidungen,

154 Hubig 2006.
155 Rip 2007.

zum Beispiel in Bezug auf die Wahl des Wohnortes oder die Ansiedlung von Betrieben. Die Akteur-Netzwerk-Theorie[156] sowie unterschiedliche Sichtweisen innerhalb der Technikphilosophie[157] haben die Gedanken von Ko-Konstruktion und Ko-Evolution erweitert und teils die technischen Objekte aufgrund ihres Einflusses auf den Menschen als Ko-Akteure (Aktanten) definiert.

4.3 Mehrstufige Mensch-Technik-Wechselwirkungen

Die zunehmende Komplexität der Mensch-Technik- bzw. Mensch-Maschine-Relation verändert auch deren Wahrnehmung. Insbesondere KI-gesteuerte Systeme wie Produktionsroboter, „autonome" Fahrzeuge, Therapieprogramme oder Schachcomputer sind Beispiele digitaler Technik, in denen die vormals klaren Unterscheidungen von Mensch und Technik weniger eindeutig zu werden scheinen. Androide Roboter erscheinen menschenähnlich, Hilfesuchende interagieren mit Therapieprogrammen, als ob es sich um menschliche Therapiekräfte handeln würde, und der Schachcomputer scheint die Partie gewinnen „zu wollen". Die Anthropomorphisierung digitaler Technik ist in der Umgangssprache weit fortgeschritten. Sie zeigt sich darin, dass KI und Robotern Fähigkeiten wie Denken, Lernen, Entscheiden oder Emotionalität zugeschrieben werden, wodurch sie scheinbar in die Gemeinschaft der denkenden, lernenden, entscheidenden und fühlenden Menschen aufgenommen werden.

Phänomenologisch geht damit einher, dass sich durch „autonom" werdende KI-gestützte Technik Subjekt-Objekt-Verhältnisse zwischen Mensch und Technik verändern. Im traditionellen Bild gestalten und nutzen menschliche Subjekte

156 Latour 2007.
157 Hubig 2006; Ihde 1990.

technische Objekte. Bereits mit traditioneller Software, mehr noch mit KI, kommt es jedoch zu neuen Konstellationen. In vernetzten Systemen haben Menschen teils die Subjekt-, teils aber auch die Objektrolle inne. Wenn einerseits Entscheidungen über Menschen an Softwaresysteme delegiert werden, beispielsweise hinsichtlich der Gewährung von Krediten oder Sozialleistungen, werden Menschen zu Objekten der „Entscheidungen" dieser Systeme, die hier auftreten, als ob sie Subjekte seien. Andererseits kann die Subjektrolle von Menschen durch gute Software zur Entscheidungsunterstützung erhöht werden, beispielsweise wenn diese qualitativ hochwertige, diskriminierungsfreie und nachvollziehbare Informationen liefern, welche die Qualität menschlicher Entscheidungen und deren Begründbarkeit verbessern. Verschiebungen in den Subjekt-Objekt-Rollen zwischen Mensch und Technik müssen daher differenziert betrachtet werden. Sie hängen einerseits vom Ausmaß und diversen technischen und organisationalen Details ab; andererseits – und dies ist von besonderer ethischer Relevanz – manifestieren sie sich bei verschiedenen Personengruppen auf unterschiedliche Weise.

Die Gestaltung der Software und der damit operierenden Maschinen gibt jeweils die Alternativen vor, innerhalb derer Menschen handeln können. Optionen, die in dem Design nicht vorgesehen sind, werden ausgeschlossen. Algorithmen und Maschinen regulieren somit menschliches Handeln.[158] Derartige Prozesse finden in der Digitalisierung seit Jahrzehnten statt, werden aber durch KI-Systeme verschärft. Menschliche Akteure erleben dadurch einerseits eine Verminderung ihrer Autorschaft über das eigene Handeln und fühlen sich zunehmend eingeschränkt und fremdbestimmt. Andererseits werden KI-Systeme zielgerichtet von Akteuren eingesetzt, um die eigenen Handlungsmöglichkeiten zu erweitern. Ein Beispiel hierfür sind ADM-Systeme, die Klassifikationen und Prognosen vornehmen und beispielsweise durch das Errechnen von

158 Orwat et al. 2010.

Risikoscores Menschen bei der Entscheidungsfindung unterstützen (vgl. Beispiele in Teil II). Es ist immer wieder ein erwünschter Effekt erweiterter Autorschaft, wenn menschliche Entscheidungen dadurch auf eine sachlichere Grundlage gestellt werden. Auf der anderen Seite jedoch droht das Risiko, dass Menschen den Ergebnissen der KI-Systeme, auch wenn diese nur als Vorschläge unterbreitet werden, einfach blind folgen. Dann würde die Person, die eine Entscheidung trifft, eher *reagieren* als aus eigener Einsicht heraus *agieren*, was ihre Autorschaft vermindern würde (Automation Bias). Neben den bereits vielfach diskutierten Herausforderungen an eine transparente und rechtssichere Zuschreibung von Verantwortung in komplexen Mensch-Technik-Systemen, etwa beim „autonomen" Fahren[159], ist von anthropologischer und ethischer Relevanz, inwieweit die digitalen Systeme Menschen unterstützen und dadurch die Möglichkeit der Entfaltung der menschlichen Fähigkeiten erweitern oder durch technische, von den Herstellern der Systeme vorgegebene Handlungsschemata diese Entfaltung behindern und vermindern.

Die angedeutete Komplexität neuer Mensch-Maschine-Wechselwirkungen und die erwähnten Verschiebungen in Subjekt-Objekt-Rollen haben nicht nur in der öffentlichen Debatte, sondern auch in der Wissenschaft zu einer Aufweichung eines strikten Dualismus zwischen Mensch und Maschine geführt. So wird vor allem in der Wissenschafts- und Technikforschung (Science and Technology Studies) sowie der Techniksoziologie seit Jahrzehnten über die „Handlungsträgerschaft" von technischen Systemen diskutiert. Einige dieser Positionen, insbesondere die Akteur-Netzwerk-Theorie[160], die einen sehr schwachen Handlungsbegriff propagiert und diesen auf viele Entitäten ausweitet, stehen dabei in deutlicher Spannung mit der im vorigen Kapitel ausführten philosophischen Handlungstheorie, die einen anspruchsvollen Handlungsbegriff

159 Ethik-Kommission Automatisiertes und vernetztes Fahren 2017.
160 Latour 2007; Law/Hassard 1999.

beschreibt und diesen auf menschliche Akteure beschränkt. Jenseits disziplinärer Einzelperspektiven stellen sich in Anbetracht der zunehmenden Verschränkung und wechselseitigen Beeinflussung von Menschen und Maschinen unter anderem folgende Fragen:

- Reicht das technische Vokabular zur Beschreibung von KI-Systemen noch aus, um Phänomene komplexer Interaktionen von Mensch und Maschine zu beschreiben?
- Inwiefern kann bzw. sollte davon gesprochen werden, dass Maschinen handeln oder mithandeln können? Sollte also Maschinen die Rolle von Akteuren zugesprochen werden und wenn ja, unter welchen Bedingungen und mit welchen Implikationen?
- Kommt es zu Verschiebungen in den Möglichkeiten menschlicher Autorschaft und wenn ja, in welchen Richtungen?

Die Akteur-Netzwerk-Theorie beantwortet diese Fragen, indem sie technischen Systemen den Status von Akteuren mit eigenen Dynamiken zuspricht und von hybriden Handlungszusammenhängen zwischen Mensch und Maschine ausgeht.[161] Als Beobachtungstheorie ohne spezifische normativ-anthropologische Prämissen hilft sie, vermeintlich autonome Wirkungen von Techniken, Artefakten oder Sachen und deren gesellschaftsveränderndes Potenzial zu erkennen. So kann es gelingen, Phänomene aus der Beobachterperspektive in den Blick zu nehmen, die mit einem starken, die Autonomie betonenden Handlungsbegriff tendenziell ausgeblendet werden. Freilich bleibt die Frage nach dem Zusammenhang zwischen Handlung und Verantwortung hier offen.

Der Soziologe Werner Rammert, einer der Pioniere der Thematisierung möglicher Handlungsträgerschaft von Technik, schlägt einen Mittelweg zwischen einer anspruchsvoll

161 Latour 2007.

normativen Vorstellung von Handeln und der Vorstellung eigenmächtigen maschinellen Agierens vor. Stattdessen soll von einer verteilten Handlungsträgerschaft zwischen Mensch und Maschine gesprochen werden, um die Vorstellung zu vermeiden, Technik sei etwas außerhalb des Sozialen Stehendes.[162] So geht Rammert wie die Akteur-Netzwerk-Theorie von hybriden, soziotechnischen Konstellationen aus, in denen Menschen und Maschinen auf komplexe Weise wechselwirken. Das Handeln des Menschen in dieser Perspektive sieht er zwar von technischen Prozessen beeinflusst, jedoch nicht als determiniert an. Der Einfluss der Technik kann sich in beiden Richtungen auswirken: Individuelle Freiheitsspielräume und die Entfaltung der menschlichen Autorschaft des eigenen Lebens können sowohl erweitert als auch vermindert werden.

Vor diesem Hintergrund schlägt Rammert vor, die Wechselwirkung von Mensch und Technik dreistufig zu beschreiben, um sowohl die Komplexität dieser Wechselwirkungen empirisch zu erfassen als auch die Zuschreibung von Verantwortung auf Menschen zu begrenzen. Als Stufe 1 nennt er *Kausalität* im Sinne von, Veränderung bewirken zu können. Stufe 2 beschreibt er als *Kontingenz* mit der Bedeutung, auch anders agieren zu können. Stufe 3 schließlich ist durch *Intentionalität* gekennzeichnet, was beinhaltet, das eigene Verhalten deuten und steuern zu können. Kausalität und Kontingenz charakterisieren nach Rammert nicht nur die menschliche Intervention, sondern auch die von Technologien. Algorithmen „wählen" zwischen Alternativen; automatisierte Entscheidungssysteme „bestimmen" einen Risikoscore, auf dessen Basis entweder ein Mensch eine Entscheidung trifft, beispielsweise über eine Kreditvergabe, oder die Software „entscheidet" sogar selbst, welche Bewerbungen bereits vorab aussortiert und der Personalabteilung erst gar nicht angezeigt werden. Algorithmen wirken hier auf mannigfaltige und höchst komplexe Weise, ohne dass ihnen dafür Intentionalität unterstellt werden kann.

162 Rammert/Schulz-Schaeffer 2002.

Intentionalität nämlich, die höchste Stufe des „Agierens", ist dem Handeln von Menschen vorbehalten. Daher kann nach Rammert nur ihnen Verantwortung zugeschrieben werden.

Die genannten Schwierigkeiten bei der Zuschreibung von Verantwortung haben auch den Philosophen Luciano Floridi motiviert, eine Theorie „verteilter Moralität"[163] zu entwickeln, die auf die Komplexität digitaler Mensch-Maschine-Systeme zugeschnitten ist. Basierend auf früheren Arbeiten[164] unterscheidet er zwischen moralischer Handlungsfähigkeit (*moral agency*) und moralischer Verantwortlichkeit (*moral responsibility*).[165] Während er – wie Rammert – Intentionalität als notwendig für Verantwortlichkeit erachtet, ist sie keine Bedingung für Handlungsfähigkeit. Für Letzteres reichen Interaktivität, Autonomie und Adaptivität aus. KI-Algorithmen bzw. verteilte Mensch-KI-Systeme können demnach für ihre „Handlungen" zwar *accountable*, aber nicht *responsible* sein, da ihnen die Intentionalität fehle. Die moralischen Eigenschaften der Ergebnisse verteilter Handlungsträgerschaft *emergieren* aus den einzelnen Elementen, ohne dass ihnen eine Intention zugrunde liegt.

Auch der Versuch von Floridi, die Interaktionen von Mensch und Maschine angemessen zu beschreiben, lässt die Frage der Zuschreibung von Verantwortung offen.[166] Es gibt moralisch bedeutsame Aktionen von KI-Systemen, insofern als moralisch problematische Resultate durch KI-Systeme verursacht werden. Den Systemen kann dafür aber keine Verantwortung zugeschrieben werden. Verantwortung muss daher anders geregelt werden, zum Beispiel durch Zuschreibung von Verantwortung an relevante Institutionen bzw. Organisationen. Dies können beispielsweise die Betreiber dieser Systeme sein, da sie aufgrund ihrer intentionalen Entscheidung zu

163 Floridi 2013.
164 Floridi/Sanders 2004.
165 Da die üblichen deutschen Übersetzungen nicht immer genau zu passen scheinen, sind hier die englischen Originalbegriffe erhalten.
166 Floridi 2016.

deren Einsatz einschließlich der diffusen Verteilung von *agency* zugestimmt haben und damit – auch im Sinne von Floridi – verantwortlich sind. Der von der Europäischen Union geplante Artificial Intelligence Act wird folglich die Verantwortlichkeit konkret zuweisen müssen, um wirksam zu regulieren.

Die hier vorgestellten techniksoziologischen und -philosophischen Ansätze, die sich um ein differenziertes Verständnis der zunehmend komplexen Mensch-Technik-Wechselwirkungen in Bezug auf KI-Systeme bemühen, können mit den anthropologischen Positionen, die in Kapitel 3 dargelegt wurden, durchaus in Konflikt geraten. Entsprechend der dort ausgeführten anspruchsvollen philosophischen Handlungstheorie können Maschinen nicht handeln, weil sie nicht über Intentionen verfügen, und kommen daher als genuine Akteure nicht infrage, jedenfalls nicht nach gegenwärtigem Stand der Entwicklung. Aber auch wenn technischen Systemen keine Handlungsfähigkeit und damit Verantwortung zugeschrieben werden kann, haben sie Einfluss auf menschliches Handeln. Menschliches Handeln ist weder völlig autonom noch völlig sozial oder technisch determiniert, sondern in zunehmendem Maße soziotechnisch situiert. Auch in der Digitalisierung und der KI ist dies empirisch durch zahlreiche Studien belegt, insbesondere zum Automation Bias bzw. zu den Effekten von Nudging.[167] Diese Effekte sind höchst bedeutsam für ethische Fragen, zeigen sie doch, dass technologische Entwicklungen menschliche Handlungsfähigkeit beeinflussen und menschliche Autonomie und Autorschaft sowohl erweitern als auch vermindern können.

167 Unter Nudging versteht man die Formatierung einer Entscheidungssituation ohne an den Handlungsalternativen etwas zu verändern, sodass erwünschtes Verhalten wahrscheinlicher wird (Thaler/Sunstein 2011).

4.4 Erweiterung und Verminderung menschlicher Autonomie und Autorschaft

Angesichts der konstatierten intensiven und komplexen Wechselwirkungen zwischen Mensch und Technik stellt sich die Frage nach den Folgen des digital-technischen Fortschritts für die Bedingungen gelingenden Handelns und die Möglichkeiten, menschliche Autorschaft zu entfalten. Bisherige Erfahrungen zeigen Ambivalenzen und Dialektiken von ethischer Relevanz.[168] Neue Technologien sollen einerseits und vor allem, so die Erzählung spätestens seit Francis Bacon und der europäischen Aufklärung, die Menschen von den Zwängen der Natur und der Tradition emanzipieren, Freiheitsräume durch neue Handlungsoptionen eröffnen und damit die Entfaltungsmöglichkeiten menschlichen Handelns erweitern. Entsprechende Effekte zeigen sich in der Tat im raschen Fortschritt der digitalen Technologien: globale Kommunikation in Echtzeit, schnelle Information, Mustererkennung durch Big Data, Effizienzsteigerung und Beschleunigung der Produktion, neue Dienstleistungen und Geschäftsmodelle, bessere medizinische Diagnosen und Therapien, Roboter und Algorithmen als künstliche Assistenten, selbstfahrende Autos, Minenräumroboter und vieles mehr. Speziell die KI eröffnet Möglichkeiten, menschliches Handeln zu verbessern, so etwa durch Mustererkennung in großen Datenmengen für medizinische oder behördliche Zwecke, durch darauf aufbauende verbesserte Vorhersagen zum Beispiel zur Ausbreitung von Infektionskrankheiten oder für Prognosen in der Polizeiarbeit (Predictive Policing), durch neue Möglichkeiten individualisierter Information und Werbung, aber auch durch Anwendungen im Bildungsbereich. Technik ist zentraler Teil menschlichen Lebens und gesellschaftlicher Vollzüge zumindest in den industrialisierten Regionen der Welt geworden und hat in vielen

168 Grunwald 2022.

Fällen eindeutig positive Folgen in dem Sinne gezeitigt, dass die Möglichkeiten menschlicher Autorschaft erweitert wurden – jedenfalls für den Teil der Erdbevölkerung, der, vor allem im globalen Norden, Zugang zu ihren Vorteilen hat.

Im Rahmen der Diffusion von Technik und Innovationen in die Gesellschaft, ihrer Nutzung und Veralltäglichung kommt es jedoch häufig zu Sekundäreffekten, die als negativ wahrgenommen werden. Zu den nicht intendierten Folgen wie Umweltproblemen und sozialen Verwerfungen zählen auch Begrenzungen menschlicher Entfaltungsmöglichkeiten. In die Erweiterung menschlicher Autonomie und Autorschaft im technischen Fortschritt ist ihre simultane Verminderung oft entweder bereits eingeschrieben oder entwickelt sich empirisch kontingent. Erweiterung und Verminderung sind häufig ineinander verschränkt, betreffen jedoch meist unterschiedliche Beteiligte in unterschiedlicher Weise, so etwa diejenigen, die Entscheidungen treffen, und diejenigen, die von diesen Entscheidungen betroffen sind. Während die Facetten der Erweiterung offensiv kommuniziert werden und oft auch deutlich sichtbar sind, etwa aufgrund neuer Fähigkeiten von IT-Systemen, ist ihre Kehrseite, die nicht immer aber immer wieder auftretende simultane Verminderung menschlicher Entfaltungsmöglichkeiten, oft nicht gut erkennbar.

Verminderungen qualifizierter Handlungsformen und Entfaltungsmöglichkeiten im Rahmen der Nutzung von Technik können etwa in folgenden Richtungen auftreten:

(1) Entstehende Abhängigkeiten: Mit dem Erfolg von Technik sind moderne Gesellschaften von ihrem reibungslosen Funktionieren abhängig geworden. Dies beginnt individuell mit der Abhängigkeit von Computer und Auto und reicht gesellschaftlich bis hin zur vollständigen Abhängigkeit vom Funktionieren der Energieversorgung und der weltweiten Datenkommunikationsnetze. Die individuelle wie kollektive Abhängigkeit von digitalen Technologien, insbesondere vom Internet, durchzieht sämtliche gesellschaftlichen Prozesse: ohne Internet keine funktionierende Weltwirtschaft, keine

Finanztransaktionen, Kollaps internationaler Logistikketten, Zusammenbruch der öffentlichen wie privaten Kommunikation. Entsprechend führt jede Übertragung von Zuständigkeiten beispielsweise durch den Einsatz von Software zur Entscheidungsunterstützung zu einer gewissen Abhängigkeit. Abhängigkeit jedoch vermindert menschliche Entfaltungsmöglichkeit, da sie Sachzwänge zum Weiterbetrieb der technischen Systeme nach sich zieht und die Vulnerabilität der Gesellschaft gegenüber technischem Versagen und intendierten Störungen (z. B. Hacking) steigert.

(2) Anpassungsdruck: Technik nötigt zur Anpassung. Das ist auf der Ebene konkreter technischer Objekte wie Maschinen trivial; es müssen beispielsweise Bedienungsanleitungen beachtet werden, um die Technik sachgerecht nutzen zu können. Digitale Technik jedoch reguliert und ändert subtil menschliches Handeln und Verhalten. Softwaresysteme steuern vielfach explizit oder implizit Verhalten.[169] So strukturieren privat geführte digitale Infrastrukturen die politische Kommunikation, sortieren Suchmaschinen die Welt mithilfe der von ihnen gesetzten Filter und strukturieren Onlineplattformen Geschäftsprozesse. Dahinter steht beispielsweise die schlecht durch Daten belegbare Befürchtung, dass menschliches Denken und Handeln durch fortschreitende Anpassung an Softwaresysteme allmählich nach deren Anforderungen und Vorgaben reguliert und immer stärker im technischen Sinn normiert werden könnte. Menschliche Autorschaft würde leise und unbemerkt, sozusagen durch allmähliche Gewöhnung, unkritische Übernahme algorithmischer Vorschläge (Automation Bias) und Anpassung an technische Voreinstellungen, ausgehöhlt.

(3) Verschließen von Optionen: Immer wieder werden mit dem Eröffnen neuer Handlungsspielräume andere, bis dato etablierte Optionen abgewertet oder ganz verschlossen. In der

169 Bowker/Star 1999; Nguyen 2021.

Innovationstheorie gilt dies als „schöpferische Zerstörung".[170] Dies ist einerseits der normale Gang von Transformation und Wandel. Andererseits aber stürzen Innovationen vorhandene Anerkennungs- und Wertstrukturen durch disruptive Effekte um und ziehen Gewinner wie auch Verlierer nach sich. Von den neuen Optionen profitieren häufig andere Personen und Gruppen als die, die dann im Verschließen der traditionellen Optionen zu Verlierern des Wandels werden. Zum Verschließen von Optionen menschlicher Entfaltung durch technischen Fortschritt führen unterschiedliche Mechanismen. So werden Infrastruktursysteme häufig faktisch machtförmig, indem sie Lebensformen außerhalb dieser Systeme benachteiligen oder unmöglich machen. Beispielsweise wird mittlerweile häufig die Nutzung eines Smartphones vorausgesetzt, um an bestimmten Lebensvollzügen teilnehmen zu können. Diese Form der Verschließung von Optionen kann verschiedene Bevölkerungsgruppen unterschiedlich treffen und Gerechtigkeitsprobleme mit sich bringen, wie dies zum Beispiel unter dem Stichwort „digitale Spaltung" (*digital divide*) diskutiert wird. Ein anderer Mechanismus besteht in allmählicher Gewöhnung als Folge der oben erwähnten Anpassung. Technik, gerade die Digitaltechnik, macht vielfach das Leben angenehm und komfortabel. Sobald Routinehandlungen in Beruf oder Freizeit daran adaptiert wurden, gehört diese Technik so zum Leben, dass es ohne diese Technik oft kaum noch vorstellbar ist. Alternative Optionen verschließen sich, vermeintliche Sachzwang-Argumente erwecken den Anschein der Alternativlosigkeit, sind jedoch nur Ausdruck der schleichend eingetretenen Pfadabhängigkeit durch allmähliche Anpassung und Gewöhnung.

Diese Mechanismen, in denen Optionen menschlichen Handelns sich verschließen, können im Rahmen der Diffusion neuer Technologien in die Anwendung auftreten, ohne dass Intentionen von Akteuren dahinterstehen. Es geht nicht um eine Verminderung menschlicher Autorschaft durch bewusste

170 Schumpeter 2018, 113 ff.

Delegation vormals menschlicher Handlungsvollzüge an KI-Systeme, sondern um Effekte, die schleichend und teilweise unbewusst durch Verhaltensänderungen entstehen. Das Ersetzen als Endpunkt des Delegierens vormals menschlich ausgeübter Tätigkeiten an technische Systeme erfolgt jedoch intentional. Es betrifft Funktionen und Tätigkeiten, die technomorph beschrieben und sodann von KI-gesteuerten Systemen übernommen werden können, im Idealfall funktionsäquivalent oder „besser".[171] Motivationen, menschliche Tätigkeiten durch KI-Systeme zu ersetzen, sind zum Beispiel die Effizienzsteigerung behördlicher Funktionen, die Kostensenkung in der industriellen Produktion, die Routinisierung diagnostischer Auswertungen in der Medizin oder die Ermöglichung automatisierter Überwachung in Echtzeit.

Bezogen auf die Möglichkeit menschlicher Autorschaft stellt sich die Ersetzung menschlicher Tätigkeiten zunächst als Resultat menschlicher Entscheidungen dar. So wird beispielsweise in Unternehmen oder Behörden aus unterschiedlichen Gründen die Entscheidung getroffen, bestimmte Tätigkeiten, die zuvor von Menschen durchgeführt wurden, an maschinelle Systeme zu übertragen. Dies können beispielsweise ADM-Systeme in unterschiedlichen Anwendungen sein, etwa in der Medizin, im Sicherheitsbereich oder im Sozialwesen. Diese Übertragung ist für sich genommen ein Ausdruck der Wahrnehmung menschlicher Autorschaft in bestimmten institutionellen Kontexten und unter entsprechenden Randbedingungen. Die zentrale ethische Frage ist, ob und wie diese Übertragung die Möglichkeiten *anderer* Menschen beeinflusst, vor allem jener, über die entschieden wird. Es stellt sich hier also die Frage, wie die Delegation von Tätigkeiten an Technik die Handlungsmöglichkeit und Autorschaft der Betroffenen

171 Das Wort „besser" suggeriert, dass es Verbesserung „als solche" gebe, und ignoriert, dass „besser" semantisch grundsätzlich nur im Zusammenhang mit normativen Kriterien des „besser" sinnvoll ist – und diese Kriterien können umstritten und kontrovers sein.

beeinflusst. Dies stellt sich in unterschiedlichen Anwendungsfeldern auf je andere Weise dar.

Bereits mehrfach hat sich damit gezeigt, dass mit einer erwünschten Erweiterung der Möglichkeiten menschlicher Autorschaft oft simultan eine Verminderung verbunden ist, häufig in Bezug auf andere Aspekte und Felder von Autorschaft bzw. andere Menschen. Auf jeden Fall verschieben KI-Systeme die Möglichkeiten der Wahrnehmung menschlicher Autorschaft. Dies betrifft unterschiedliche Bevölkerungsgruppen auf unterschiedliche Weise und weist daher eine soziale Dimension mit ethischen Fragen auf. Bei der Betrachtung von Chancen und Risiken etwa von Entscheidungsunterstützungssystemen im Sicherheitsbereich ist zu berücksichtigen, *für wen* es sich hier jeweils um Chancen oder Risiken, um Erweiterungen oder Verminderungen der Autorschaft handelt. Damit sind hier Aspekte sozialer Gerechtigkeit und Macht involviert. In der Digitalisierung und speziell bei der Entwicklung und Nutzung von KI ist grundsätzlich zu fragen, wer die entsprechenden Prozesse bzw. die konkreten Softwareapplikationen und KI-Algorithmen jeweils gestaltet und ob – und wenn ja mit welcher Legitimation – sie in die Autonomie und Autorschaft derjenigen, die diese Produkte nutzen, oder weiterer Betroffener eingreifen. Auch jenseits der intentionalen Manipulation durch die Gestalter sind Effekte von Beeinflussung, Gewöhnung und Abhängigkeiten bis hin zu digitalem Mediensuchtverhalten zu beobachten, in denen offenkundig menschliche Autorschaft eingeengt wird.

Die sich hier andeutenden ethischen Herausforderungen sind mit epistemologischen Herausforderungen verbunden. Allmähliche Verschiebungen, wie etwa die erwähnten Gewöhnungsprozesse an technisch normierte Handlungsmuster (Automation Bias), sind oft nur schwer aufzudecken und empirisch zu belegen. Es besteht das Risiko verspäteter Entdeckung, zu einem Zeitpunkt, zu dem möglicherweise nur noch schwer beeinflussbare Pfadabhängigkeiten bereits eingetreten sind, schlimmstenfalls ein Point of no Return überschritten wurde.

Zwischen der hohen Relevanz dieser an die Dialektik von Herr und Knecht gemahnenden Situation und der epistemologisch schwierigen Nachweislage ist die Bewusstmachung möglicher ethisch bedenklicher Zukunftsentwicklungen dieses Typs eine Herausforderung. Denn angesichts starker Gegenwartspräferenzen vieler Akteure ist sie mit den bekannten Problemen vorsorgeorientierter Kommunikation konfrontiert. Von der einen Seite droht der Vorwurf der Übertreibung, Dramatisierung oder gar Technikfeindlichkeit, von der anderen der Vorwurf der Verharmlosung. Hohe epistemologische Unsicherheit macht vorsorgeorientierte Kommunikation anfällig für Ideologie, interessengeleitete Statements und Spekulation.

4.5 Fazit

Die ethische Analyse und Beurteilung des Einsatzes von KI-Systemen bedürfen über die begriffliche, anthropologische und handlungstheoretische Vergewisserung (vgl. Kapitel 3) hinaus eines genaueren Blicks auf die sich mit der Digitalisierung verändernden Konstellationen zwischen Mensch und Technik. Im Rahmen der philosophischen Handlungstheorie können Maschinen nicht handeln und kommen als genuine Akteure mit Verantwortung nicht infrage. Dennoch haben sie Einfluss auf menschliches Handeln, das in modernen Gesellschaften in zunehmendem Maße soziotechnisch situiert ist. Erfahrung und empirische Forschung zeigen, dass Technik einerseits von Menschen als Mittel nach Zwecken gestaltet wird, dass aber andererseits neue Technik und darauf aufbauende Innovationen oder Dienstleistungen menschliches Handeln und Verhalten beeinflussen. Ethisch relevant ist insbesondere, wie sich diese Wechselwirkungen auf die Möglichkeiten menschlicher Autorschaft und Verantwortungsübernahme auswirken und wie diese sich angesichts der zunehmenden Verbreitung von KI-Systemen verändern.

KI-Systeme können in vielen Feldern menschliche Handlungen und Entscheidungen unterstützen, dadurch zu besseren Ergebnissen beitragen und damit menschliche Autorschaft erweitern. Vor allem die durch Algorithmen eröffnete Möglichkeit, in großen Datenmengen (Big Data) Muster zu erkennen, die den Menschen ansonsten verborgen wären, ist die Basis für den unterstützenden Einsatz der KI zum Beispiel in der medizinischen Diagnostik, im Bildungsbereich, aber auch im Medienbereich und in der Verwaltung. Gerade im Bereich der sozialen Medien zeigt sich hier ein Phänomen, das als Hypernudge[172] beschrieben wurde: Datenbasierte algorithmische Systeme kuratieren dynamisch hochgradig personalisierte Informationsumgebungen, denen man sich nur schwer entziehen kann. Menschliche Autorschaft kann also durch KI nicht nur erweitert, sondern auch vermindert werden, entweder durch intendierte Delegation von Entscheidungen an automatische Systeme oder durch allmähliche Effekte der Gewöhnung und Anpassung an datengenerierte Empfehlungen von KI-Systemen.

KI-Systeme verschieben die Möglichkeiten der Wahrnehmung menschlicher Autorschaft. Die grundsätzlich erwünschte Erweiterung von Autorschaft ist häufig simultan mit einer Verminderung in Bezug auf andere Aspekte von Autorschaft bzw. andere Akteure verbunden. Insbesondere sind verschiedene Akteursgruppen in unterschiedlicher Weise betroffen. Die ethische Analyse von Chancen und Risiken etwa von ADM-Systemen in der öffentlichen Verwaltung (vgl. Kapitel 8) muss darauf achten, *für wen* es zu Erweiterungen oder Verminderungen der Autorschaft kommt, etwa in der Differenz von Entscheidenden und Betroffenen. Es sind also mit dem Einsatz von KI-Systemen auch Fragen von Gerechtigkeit und Autonomie- bzw. Machtverteilung involviert. Speziell ist zu fragen, ob und wie die jeweiligen Gestalter der entsprechenden Prozesse bzw. der konkreten Softwareapplikationen und

172 Yeung 2017.

KI-Algorithmen in die Autonomie und Autorschaft derjenigen eingreifen, die diese Produkte nutzen oder anderweitig von ihnen betroffen sind.

Weiterhin sind psychologische Effekte zu beachten, die spezifisch für digitale Instrumente und insbesondere für KI-Systeme sind. Hier ist vor allem der Automation Bias zu nennen. Menschen vertrauen, so empirische Untersuchungen, algorithmisch erzeugten Ergebnissen und automatisierten Entscheidungsprozeduren häufig mehr als menschlichen Entscheidungen. Vermutlich spielen dabei verbreitete Objektivitätsunterstellungen gegenüber Daten und Rechenverfahren eine Rolle, während menschliche Urteile tendenziell als subjektiv wahrgenommen werden. Gerade bei Entscheidungen, die mit einer großen prognostischen Unsicherheit konfrontiert sind und zugleich gravierende Auswirkungen haben, besteht die latente Tendenz, den datenbasierten algorithmischen Auswertungen mehr zu vertrauen. Damit wird Verantwortung – zumindest unbewusst – auf diese „Quasi-Akteure" delegiert. Dieser Bias zugunsten der algorithmischen Verfahren kann beispielsweise dazu führen, dass auch bei einer handlungstheoretisch korrekten Organisation von Entscheidungsprozessen, in denen ein KI-System normativ strikt auf die Rolle der Entscheidungsunterstützung begrenzt und dem Mensch, der die Entscheidung trifft, die Verantwortung zugeschrieben wird, das KI-System allmählich in die Rolle des eigentlichen „Entscheiders" gerät und menschliche Autorschaft und Verantwortung ausgehöhlt werden. Bisweilen wird versucht, dieser Gefahr vorzubeugen, indem bei Verwendung eines Entscheidungsunterstützungstools ein entsprechender Warnhinweis gegeben wird. Eine weitere denkbare Vorkehrung wäre die Verpflichtung der entscheidenden Fachkräfte, die etwaige Übernahme des algorithmischen Entscheidungsvorschlages – etwa mit Verweis auf die eigene erfahrungsbezogen intuitive oder kollegial erörterte Prognose – ausdrücklich zu begründen. Auf jeden Fall bedarf dieser Überlappungsbereich normativer Regulierung und empirisch-psychologischer Effekte

besonderer Aufmerksamkeit in den ethischen Analysen zu den Anwendungsfeldern in den folgenden Kapiteln.

Menschen, die Entscheidungen treffen, haben kaum eine Möglichkeit, die epistemische Evidenz der Korrelationen und Muster kritisch zu beurteilen, sondern sind vielfach darauf angewiesen, sie so zu nehmen, wie sie von den Systemen bereitgestellt werden. Sie unterliegen damit einem verborgenen Nudging durch die Art und Weise, wie die KI-Systeme zu ihren Ergebnissen kommen, und werden in bestimmte Richtungen des Entscheidens gedrängt. Mögliche Einseitigkeiten, etwa auf Basis der Datenlage, sowie daraus möglicherweise resultierende Diskriminierungen geraten aus dem Blick und menschliche Autorschaft wird entleert.

In den Abwägungen zwischen Erweiterung und Verminderung menschlicher Autorschaft in ihrer sozialen Verteilung sind bereits auf einer abstrakten ethischen Ebene mehrere Dimensionen zu berücksichtigen. Erstens bedarf die Übertragung menschlicher Tätigkeiten auf KI-Systeme der Transparenz gegenüber den Betroffenen. Sie sollten darüber informiert sein, auf welche Weise Entscheidungen zustande kommen, von denen sie dann betroffen sind. Dies hat zweitens mit der Klarstellung von Verantwortungszuschreibungen zu tun. Um ein Verantwortungs- und gegebenenfalls auch Haftungsvakuum zu verhindern, muss die Verantwortungszuschreibung etwa über die Betreiber der Systeme oder die menschlichen Akteure, die die Übertragung an sie beschlossen haben, geregelt werden. Drittens bedarf es der Sicherstellung der Nachvollziehbarkeit in Bezug auf das zweckhafte Funktionieren der KI-Systeme. Viertens müssen mögliche nicht intendierte Folgen wie beispielsweise schleichend einkehrende Abhängigkeiten von den KI-Systemen oder allmähliche Aushöhlung menschlicher Autorschaft sorgfältig beobachtet werden, um gegebenenfalls rechtzeitig korrigierend eingreifen zu können.

TEIL II:
AUSGEWÄHLTE ANWENDUNGEN UND SEKTORSPEZIFISCHE EMPFEHLUNGEN

In den folgenden Kapiteln des zweiten Teils dieser Stellungnahme sollen die zuvor angestellten Überlegungen anhand von Analysen in vier verschiedenen Sektoren vertieft werden: der Medizin (Kapitel 5), der (schulischen) Bildung (Kapitel 6), der öffentlichen Kommunikation und Meinungsbildung (Kapitel 7) sowie der öffentlichen Verwaltung (Kapitel 8).

KI-basierte Softwaresysteme werden in allen vier Sektoren eingesetzt, wobei sich die Durchdringungsbreite und -tiefe durchaus unterscheiden. So liefern die für den Bereich der öffentlichen Kommunikation und Meinungsbildung zentralen sozialen Medien ein Paradebeispiel einer sehr umfassenden Delegation vormals menschlicher Tätigkeiten an Algorithmen, beispielsweise zu Zwecken der Kuratierung und Moderation von Inhalten. Der Bereich der schulischen Bildung bildet den anderen Pol, gibt es hier doch eine vergleichsweise geringe Nutzung digitaler Technologien im Allgemeinen und von KI-Systemen im Speziellen, eine vollständige Ersetzung menschlicher Tätigkeiten scheint in weiter Ferne. Im Bereich der Medizin hingegen werden zunehmend einzelne Tätigkeiten oder gar ganze Funktionen an KI-basierte Softwaresysteme delegiert. Dies reicht vom Einsatz von KI-basierter Mustererkennung zu Zwecken der Krebsdiagnostik bis hin zur Verwendung von Chatbots in der Therapie, die schon einen möglichen Ersatz menschlichen Fachpersonals suggerieren. Auch in die öffentliche Verwaltung haben KI-basierte Softwaresysteme Einzug gehalten, insbesondere in Form von datenbasierter Software, die zur Erstellung von Risikoprofilen und zur Entscheidungsunterstützung herangezogen wird.

Die folgenden Analysen werden zeigen, dass Fragen des richtigen Ausmaßes der Delegation von Tätigkeiten und Funktionen an Softwaresysteme nur kontext-, anwendungs- und personenbezogen spezifiziert werden können. Dabei gilt als Richtschnur der Bewertung, ob die Delegation zu einer Erweiterung der Handlungsmöglichkeiten, insbesondere der Möglichkeiten für verantwortliches Handeln und Autorschaft, der verschiedenen betroffenen Akteure führt oder ob

es in der Folge möglicherweise zu einer Verminderung solcher Handlungsmöglichkeiten sowie zu negativen Auswirkungen auf Möglichkeiten der Autorschaft und Verantwortungsübernahme kommt. Die Analyse des Einsatzes von KI-Systemen in den vier Sektoren endet jeweils in sektorspezifischen Empfehlungen. Unabhängig davon lassen sich aus dem Vergleich der Sektoren auch übergreifende Themen ausmachen, Aspekte, die sich in den vier Sektoren mal ähnlich, mal aber sehr unterschiedlich darstellen. Diese Themen werden in Teil III „Querschnittsthemen und übergreifende Empfehlungen" aufgegriffen.

5 MEDIZIN

5.1 Einleitung

Die zunehmende Durchdringung unserer Lebenswelt mit digitalen Produkten, die über KI-Komponenten verfügen, breitet sich auch im Gesundheitssystem immer weiter aus. Angesichts der Dynamik der technischen Entwicklung und der ökonomischen Triebkräfte dieses Transformationsprozesses ist es aus ethischer Perspektive erforderlich, die damit verbundenen Herausforderungen möglichst präzise wahrzunehmen, die jeweiligen Vor- und Nachteile des Einsatzes von KI-Instrumenten differenziert abzuwägen und einer problematischen Verselbstständigungstendenz des Technikeinsatzes frühzeitig entgegenzuwirken. Zwar gibt es bereits eine Vielzahl wertvoller ethischer Empfehlungen und Richtlinien, die sich auf unterschiedlichen Ebenen mit den Bedingungen eines verantwortbaren KI-Einsatzes beschäftigen[173], doch wird dabei entweder ein sehr genereller Überblick über mehrere Handlungsfelder gegeben, um die Chancen und Risiken der Schaffung und Bewirtschaftung integrierter Datenräume mit neuen KI-Technologien auszuloten, oder es werden die einschlägigen Herausforderungen im Gesundheitsbereich nur aus einer begrenzten – etwa ärztlichen – Perspektive reflektiert, um die Handlungssicherheit einer bestimmten Personengruppe im Umgang mit den neuen Instrumenten zu verbessern.[174]

Die vorliegende Stellungnahme knüpft an diese Vorarbeiten an, indem sie nicht nur die vielfach erhobene Forderung nach einem „Vorrang menschlichen Handelns" mit Blick auf die verschiedenen Formen der Mensch-Maschine-Relation auf diesem besonders sensiblen Handlungsfeld näher spezifiziert

173 Europäische Kommission 2020; Hochrangige Expertengruppe für künstliche Intelligenz 2019.
174 Zentrale Ethikkommission bei der Bundesärztekammer 2021.

und in Bezug setzt zu den bekannten sektorübergreifenden Herausforderungen (z. B. des Datenschutzes, der Qualität, Robustheit und Sicherheit der eingesetzten Instrumente angesichts verschiedener Manipulationsmöglichkeiten, der Bias- bzw. Diskriminierungsgefahren infolge einseitiger Trainingsdaten, der Nachvollziehbarkeits- bzw. Opazitätsprobleme besonders komplexer Algorithmen – siehe Kapitel 2).

Aufgrund der Komplexität des medizinischen Versorgungssystems bedarf es auch mit Blick auf die folgende genauere Betrachtung von KI-Anwendungen im Gesundheitswesen einer wenigstens dreifachen Differenzierung:

Erstens sind mehrere Akteursgruppen zu unterscheiden, die bezüglich eines KI-Einsatzes unterschiedliche Funktionen und Verantwortlichkeiten besitzen, wie zum Beispiel Personen in der Softwareentwicklung, Kontroll- und Zertifizierungsinstanzen, medizinische Forschungsgruppen, klinisch tätige Personen, die von ihnen versorgten Menschen, medizinische Fachgesellschaften sowie Träger von medizinischen Versorgungseinrichtungen.

Zweitens umfasst das Gesundheitswesen mit der Grundlagen- und (prä-)klinischen Forschung sowie der konkreten medizinischen Versorgung durch Prävention, Kuration und Palliation, denen wiederum verschiedene diagnostische, therapeutische und prognostische Einzelmaßnahmen zuzuordnen sind, ganz unterschiedliche Anwendungsbereiche für KI-Produkte, deren jeweilige Chancen und Risiken kontextsensitiv zu beurteilen sind.[175]

Drittens sind hinsichtlich der derzeit verfügbaren KI-basierten Instrumente unterschiedliche *Grade der Ersetzung* menschlicher Handlungssegmente zu beobachten, die im Kernbereich medizinischer Behandlungen ganz verschiedene Auswirkungen auf die Arzt-Patienten-Beziehung haben können und dadurch die Handlungsmöglichkeiten der verschiedenen

175 Der Bereich der privaten Nutzung entsprechender Produkte im Rahmen einer *digital self-care* bleibt hier ausgeklammert.

Akteure sowohl erweitern als auch vermindern können.[176] Das Spektrum der Einsatzmöglichkeiten dieser Instrumente reicht dabei von einer engen Nutzung, bei der lediglich ein einzelnes Segment ärztlichen Handelns technisch substituiert wird, über komplexere KI-Anwendungen, die mehrere ärztliche Handlungsschritte begleiten und unterstützen, bis hin zu einer vollständigen Ersetzung des Behandlers durch ein KI-System. Letztere trägt zwar gegenwärtig noch weithin fiktionale Züge, ist aber in einzelnen Sektoren der medizinischen Versorgung bereits Realität geworden.

Nachfolgend werden anhand exemplarischer Beispiele die jeweiligen Chancen und Risiken eines KI-Einsatzes in unterschiedlichen Bereichen des medizinischen Systems analysiert.

5.2 Einsatz von KI-Systemen in der Medizin

Eine verantwortliche Verwendung von KI-Systemen im Bereich der Medizin setzt voraus, dass die gesamte Handlungskette – von der Entwicklung entsprechender Produkte über ihren Einsatz in der Forschung bis hin zu ihrer Implementierung in den verschiedenen Sektoren der medizinischen Versorgung – ethischen Standards genügt, kontinuierlich überwacht und gezielt so weiterentwickelt wird, dass Vorteile sukzessive immer besser genutzt und Gefahren vermieden werden.

5.2.1 Entwicklung von KI-Systemen

Bereits die Entwicklung geeigneter KI-Komponenten für die medizinische Praxis, die hier als erster, dem klinischen Einsatz solcher Systeme notwendig vorausgehender, ethisch relevanter

176 Zum Einsatz KI-gestützter robotischer Systeme im Rahmen der Pflege vgl. Deutscher Ethikrat 2020a.

Teilbereich erwähnt werden soll, stellt eine anspruchsvolle Aufgabe dar. Sie erfordert nicht nur eine enge interdisziplinäre Zusammenarbeit verschiedener Sachverständiger aus unterschiedlichen Fachgebieten (z. B. Informatik, Ingenieurwissenschaft, Medizin, Recht), um geeignete Algorithmen für die zuverlässige Bewältigung bestimmter Aufgaben definieren zu können, sondern stellt auch hohe Anforderungen an die Qualität der verwendeten Trainingsdaten, um vermeidbare Verzerrungen der Ergebnisse von vornherein auszuschließen oder zumindest zu minimieren.

Eine Besonderheit von KI-Systemen, die auf mit maschinellem Lernen aus Daten gewonnenen Modellen basieren, besteht darin, dass bei manchen Systemen selbst diejenigen, die diese Instrumente entwickeln, aufgrund der enormen Komplexität der Datenverarbeitungsprozesse nicht mehr rekonstruieren können, wie bestimmte Resultate zustande gekommen sind, da die Eingaben mit hochgradig nichtlinearen und verteilten Prozessen verarbeitet werden. Der Prozess wird zur Blackbox. Zwar kann eine solche Opazität auch darin ihren Ursprung haben, dass bestimmte Algorithmen urheberrechtlich geschützt sind, doch gibt es neben dieser von äußeren Faktoren bedingten und insofern kontingenten Unzugänglichkeit auch eine rein technisch bedingte Opazität, die sich je nach Kontext unterschiedlich auswirken kann. Während es in manchen Bereichen sinnvoll oder sogar notwendig sein kann, ein Höchstmaß an Erklärbarkeit der jeweiligen Resultate anzustreben (Explainable AI), wofür gegebenenfalls ein hoher technischer und finanzieller Aufwand erforderlich ist, dürfte es in anderen Bereichen ausreichen sicherzustellen, dass die Personen, die diese Systeme anwenden, deren Resultate stets einer eigenen Plausibilitätsprüfung unterziehen, um den Gefahren eines ungerechtfertigten blinden Vertrauens in die Technik (Automation Bias) zu entgehen.

Die Forderung nach Transparenz lässt verschiedene Grade zu, sodass je nach Anwendungsbereich zu prüfen ist, was aus welchen Gründen für wen in welchem Umfang und zu

welchem Zweck erklärbar sein muss. Tatsächlich gibt es in der Medizin viele Beispiele nicht nur für den Einsatz technischer Geräte, deren genaue Wirkmechanismen denjenigen, die sie in ihrer Berufspraxis anwenden, allenfalls ansatzweise durchsichtig sind, sondern auch dafür, dass bestimmte Interventionen auch dann sinnvoll geplant und durchgeführt werden können, wenn das kausale Wissen um konkrete Wirkmechanismen begrenzt ist. Angesichts des Technisierungsgrades der modernen Medizin ist es weder möglich noch erforderlich, dass die behandelnden Personen die internen Prozesse der von ihnen genutzten technischen Hilfsmittel stets im Detail durchschauen, solange diese Prozesse in ausreichendem Umfang zumindest durch geeignete Stellen nachvollzogen und damit überprüft werden können.

In derart gelagerten Fällen ist es umso wichtiger, „mittels geeigneter Prüf-, Zertifizier- und Auditierungsmaßnahmen sicherzustellen“[177], dass die jeweiligen Systeme technisch einwandfrei funktionieren und verantwortungsvoll eingesetzt werden. Die Zertifizierung vertrauenswürdiger KI-Systeme erstreckt sich insbesondere auf die Wahrung von Mindest- und anwendungsbezogenen Spezialanforderungen bezüglich der Autonomie und Kontrolle, der Fairness, der Transparenz, der Verlässlichkeit, der Sicherheit und des Datenschutzes.[178]

Sowohl die medizinischen Fachleute, die solche „intelligenten“ Medizinprodukte verwenden, als auch die von ihnen behandelten Personen müssen darauf vertrauen können, dass nur hinreichend geprüfte und nach möglichst international konsentierten Maßstäben zertifizierte KI-Produkte zum Einsatz kommen und von der jeweiligen Gesundheitseinrichtung vorschriftmäßig gewartet und gegen Manipulation geschützt werden. Dies gilt umso mehr, als die Lebensdauer von Softwareprodukten oft relativ kurz ist und die Systeme

177 Zentrale Ethikkommission bei der Bundesärztekammer 2021, A5.
178 Vgl. beispielsweise das Whitepaper des Projekts „Zertifizierte KI“ der Kompetenzplattform KI.NRW (Cremers et al. 2019).

durch häufige Updates einen insgesamt fließenden Charakter aufweisen.

Angesichts der Dynamik der technischen Entwicklung und der wirtschaftlichen Interessen derjenigen – nicht selten als Start-up aus universitären Forschungseinrichtungen heraus gegründeten – Unternehmen, die solche Produkte entwickeln und vertreiben, bedarf es eines geordneten Zusammenwirkens von staatlichen Aufsichts- und Kontrollbehörden, der Entwicklung neuer bzw. ergänzter Leitlinien für deren Einsatz durch die jeweils zuständigen medizinischen Fachgesellschaften sowie kontinuierlicher Anstrengungen zur weiteren Aus-, Fort- und Weiterbildung des medizinischen Personals, um einen verantwortlichen Umgang mit den neuen Techniken zu gewährleisten.

Von besonderer ethischer Brisanz sind jene KI-Systeme, bei denen selbst von den Personen, die das System entwickeln und programmieren, nicht mehr vollständig nachvollzogen werden kann, wie ein bestimmtes Ergebnis zustande kommt. Zumindest insofern KI-Systeme medizinische Entscheidungsvorschläge mit schwerwiegenden Konsequenzen für das Überleben und grundlegende Aspekte der Lebensqualität unterbreiten, müssen deren grundlegende Funktionsweisen und Arbeitsprozesse erklär- sowie interpretierbar sein, um einen selbstbestimmten Einsatz zu gewährleisten. Lassen sich die anwendungsbezogenen Transparenzanforderungen – selbst unter Rückgriff auf Erkenntnisse der Explainable-AI-Forschung – nicht erfüllen, verbietet sich jedenfalls ein Einsatz in diesem Bereich der medizinischen Versorgung. Die Festlegung derartiger Bereiche erfordert einen breiten gesellschaftlichen Diskurs, in den insbesondere die sektorbezogenen Chancen und Risiken des Einsatzes intransparenter KI-Systeme einzubeziehen sind.

5.2.2 KI in der medizinischen Forschung

Da die Planung und Durchführung medizinischer Forschung besonders hohe Anforderungen an den reflektierten Umgang mit unterschiedlichen Wissensformen und die methodischen Fertigkeiten derer stellt, die konkrete Forschungsfragen entwickeln und sie durch geeignete Experimente auf empirisch überprüfbare Weise zu beantworten suchen, sind die in dieser Weise am Forschungsprozess beteiligten Personen grundsätzlich nicht durch KI-Systeme ersetzbar. Trotz dieser prinzipiellen Grenzen kann die gezielte Implementierung von KI-Elementen auch im Kontext der medizinischen Forschung in mehrfacher Hinsicht vorteilhaft sein, sofern der Schutz derjenigen, die an den Studien teilnehmen, und ihrer personenbezogenen (Gesundheits-)Daten gewährleistet ist. Dazu bedarf es gerade in kollaborativen Forschungsprojekten nicht nur präziser Bestimmungen, „wer unter welchen Bedingungen Zugang zu den Daten erhält, um einen transparenten Austausch bei gleichzeitiger Gewährleistung hoher Datenschutzstandards zu ermöglichen", sondern auch „klarer Vorgaben, unter welchen Umständen Probanden und Patienten Zugang zu ihren Daten haben können, wie sie modular und dynamisch in ihre Nutzung einwilligen können […] und wie die Daten langfristig erhalten bleiben".[179]

Das Spektrum eines sinnvollen KI-Einsatzes in der medizinischen Forschung ist weit: Neben hilfreichen Vor- und Zuarbeiten etwa bei Literaturrecherchen oder der Durchsuchung großer Datenbanken können KI-Instrumente einen Beitrag dazu leisten, neue Korrelationen zwischen bestimmten Phänomenen zu entdecken, deren Bedeutung für die medizinische Praxis jedoch durch gezielte Analysen überprüft werden muss. Drei Beispiele mögen dies verdeutlichen:

Das erste Beispiel betrifft den Einsatz von KI bei der Vorhersage raum- und zeitabhängiger medizinrelevanter Prozesse,

179 Deutscher Ethikrat 2017, 97.

wie etwa der Ausbreitung eines Virus, in der Epidemiologie. In diesen Fällen ist die Anwendung mathematischer Algorithmen[180] häufig die einzig effiziente Methode, um prospektiv zu relevanten Erkenntnissen zu gelangen. Da man meist nicht alle relevanten Parameter kennt, sind die Vorhersagen mit einem Unsicherheitsfaktor verbunden, der – ähnlich einer Wettervorhersage – umso größer wird, je weiter die Vorhersage in der Zukunft liegt und je feiner die zeitliche und räumliche Auflösung ist. Es können jedoch Szenarien unter verschiedenen Annahmen berechnet werden, sodass ein Bereich wahrscheinlicher Entwicklungen sinnvoll eingegrenzt werden kann.

Als zweites Beispiel sei der KI-Einsatz in der Vorhersage komplexer Molekülstrukturen genannt, der bei der Erforschung von Krankheitsmechanismen und der Entwicklung vieler Therapeutika eine große Rolle spielt. Hier werden Algorithmen genutzt, die vor allem auf der von DeepMind (einer Tochter des Google-Mutterkonzerns Alphabet) entwickelten Software AlphaFold basieren, um beispielsweise effizient Kandidaten für Impfstoffe gegen Infektionskrankheiten (wie COVID-19, aber auch darüber hinaus[181]) zu identifizieren. Die Algorithmen sagen die Struktur von Proteinen vorher und ermöglichen so einen deutlich schnelleren Fortschritt in der Entwicklung bestimmter Substanzen, als dies über traditionelle labortechnische Verfahren möglich ist. Selbstverständlich müssen die identifizierten Therapeutika-Kandidaten weiterhin im Labor und in der Klinik getestet werden. Damit ist trotz des zeitlichen Gewinns die Sicherheit im Vergleich zu traditionellen Methoden nicht eingeschränkt.

Das dritte Beispiel stammt aus der Krebsforschung, in der die Verarbeitung großer Datenmengen etwa aus der Analyse des Genoms, des Transkriptoms und des Epigenoms von immer größerer Bedeutung wird. Einerseits bietet die

180 Dies bezieht die Modellierung über Differentialgleichungen mit ein, welche nicht in klassischem Sinne als Künstliche Intelligenz bezeichnet wird.
181 Higgins 2021.

systematische Verknüpfung und Integration unterschiedlicher Daten die Chance, Zusammenhänge zwischen einer Vielzahl von Variablen zu erfassen und im Blick auf mögliche kausale Einflussfaktoren zu analysieren, um zum Beispiel eine bessere Aufteilung großer Personenkollektive in relevante Subgruppen (Stratifizierung) zu ermöglichen – und damit letztlich gegebenenfalls eine besser auf diese Gruppen zugeschnittene Behandlung. Andererseits stellen die gesteigerte Komplexität und Heterogenität der Datenbasis neue Anforderungen an die methodische Kompetenz der Forschenden. Auch in der präklinischen und klinischen Forschung steht und fällt der mögliche Nutzen von KI-Anwendungen daher „mit der Expertise und Integrität der Personen oder Institutionen, die Daten generieren, auswählen, verknüpfen und interpretieren".[182]

Zwar hat die automatisierte Bearbeitung großer Datenmengen durch KI-Komponenten aufgrund ihrer gesteigerten Rechenleistung das Potenzial, „Korrelationen zwischen wesentlich mehr Faktoren schneller und besser zu entdecken und dabei auch neue Hypothesen über Wirkzusammenhänge zu entwickeln"[183], doch wäre es „ein Missverständnis zu glauben, dass mehr Daten auch *automatisch* zu mehr Wissen über kausale Effekte führen"[184]. Aufgrund der Differenz zwischen Korrelation und Kausalität bedürfen Ergebnisse maschineller Datenanalyse daher stets der unabhängigen Überprüfung und Validierung, um in der Fülle der gefundenen Korrelationen die jeweils relevanten Kausaleffekte zu identifizieren und damit den Umfang des therapierelevanten Kausalwissens zu erweitern.

Trotz aller gebotenen Vorsicht gegenüber manchen allzu euphorischen Heilsversprechen einer KI-gestützten Forschung bietet gerade der Bereich der Onkologie inzwischen reichhaltiges Anschauungsmaterial dafür, wie in der Forschung

182 Deutscher Ethikrat 2017, 67.
183 Ebd., 71.
184 Ebd., 70.

entwickelte KI-Instrumente sukzessive auch in der medizinischen Versorgung von Patientinnen und Patienten zur Diagnostik und Therapie eingesetzt werden können; dies zeigen auch einige der nachfolgend vorgestellten Beispiele.

5.2.3 KI in der medizinischen Versorgung

Obwohl sich viele der KI-basierten Systeme momentan noch in der Entwicklungs- und Erprobungsphase befinden, gibt es in den verschiedenen Bereichen der medizinischen Versorgung bereits erste Erfahrungen mit dem Einsatz dieser neuen Instrumente, die sich unter anderem durch den Grad der technischen Ersetzung einzelner oder mehrerer Handlungssequenzen voneinander unterscheiden lassen.

Enge Ersetzung

Grundsätzlich können KI-Systeme in allen Segmenten medizinischer Versorgung eingesetzt werden. Die größte praktische Verbreitung dürften derzeit Entscheidungsunterstützungssysteme in der Diagnostik haben, die versuchen, mittels computergestützter Analyse verschiedener Parameter der Labordiagnostik, der Bildbearbeitung sowie der automatisierten Durchsicht von Patientenakten und wissenschaftlichen Datenbanken Entscheidungsprozesse zu modellieren und zu automatisieren. Da die Entwicklung und Nutzung von Entscheidungsbäumen dem medizinischen Personal schon seit Langem dabei hilft, „diagnostische Informationen vollständig zu erheben, Therapieentscheidungen präzise zu treffen und damit insgesamt die bestmögliche Behandlung sicherzustellen", scheint dieses algorithmische Vorgehen geradezu „prädestiniert für KI-Anwendungen"[185] zu sein. Ihr Einsatz steht daher in struktureller Kontinuität zur Nutzung früherer

185 Schlemmer/Hohenfellner 2021, 318.

technischer Hilfsmittel der Entscheidungsfindung (z. B. bestimmter Expertensysteme).

Neu ist der Umstand, dass die KI-basierten Systeme aufgrund der gesteigerten Rechenkapazitäten sehr große Datenmengen und eine Vielzahl relevanter Parameter berücksichtigen können und die einzelnen Rechenoperationen den Anwenderinnen und Anwendern nur begrenzt zugänglich sind. So bieten vor allem Fortschritte in der Bilderkennung große diagnostische Potenziale, da durch sie zum Beispiel neue Möglichkeiten einer frühzeitigen Detektion, Lokalisation und Charakterisierung verschiedener pathologischer Veränderungen in der Gewebestruktur eröffnet werden. Das Spektrum der derzeitigen Nutzung reicht von der Untersuchung des Augenhintergrundes über die Analyse von Hautläsionen in der Dermatologie bis hin zu verschiedenen Bereichen der Onkologie, die insofern von besonderer medizinischer Bedeutung sind, als die Früherkennung maligner Tumoren essenziell für eine erfolgreiche Therapie ist. Ein Beispiel bildet hier der KI-Einsatz in der Brustkrebsdiagnostik (Infokasten 2).

Infokasten 2: Enge Ersetzung am Beispiel Brustkrebsdiagnostik

Brustkrebs (Mammakarzinom) ist mit jährlich rund 71 400 Neuerkrankungen die häufigste Krebserkrankung der Frau; nur rund ein Prozent der Brustkrebserkrankungen treten beim Mann auf.[186] Neben den vielen Einzelschicksalen, die hinter diesen Zahlen stehen, ist die kollektive Betrachtung des Brustkrebses wegen seiner Häufigkeit sowie der hohen Therapiekosten, der resultierenden Arbeitsausfälle sowie schlechter Heilungschancen bei der Erkennung des Tumors in einem späten Stadium von allgemeiner gesellschaftlicher Relevanz und sozioökonomischer Bedeutung. Eine frühzeitige Erkennung und die richtige Diagnose sind entscheidende Faktoren, um die Überlebenschancen zu erhöhen.

Um die Früherkennung zu unterstützen, wurden in vielen Ländern, so auch in Deutschland, Mammografie-Screening-Programme implementiert. Für die Mammografie werden Röntgenstrahlen eingesetzt, die mittels digitaler Detektoren nachgewiesen werden. Diese Information wird zur Rekonstruktion zweidimensionaler Bilder genutzt, die direkt für die Auswertung zur

186 https://www.krebsdaten.de/Krebs/DE/Content/Krebsarten/Brustkrebs/brustkrebs_node.html [03.03.2023].

Verfügung stehen. Üblicherweise übernimmt dies erfahrenes, radiologisch qualifiziertes Personal, aber schon seit Längerem wird daran gearbeitet, die fachärztliche radiologische Beurteilung mittels computergestützter Verfahren zu unterstützen. Die ersten sogenannten CADe- und CADx-Algorithmen (*computer-aided detection* – CADe, *computer-aided diagnosis* – CADx) konnten sich zwar nicht durchsetzen, da die Anzahl an falsch-positiven Ergebnissen zu hoch war.[187] Durch den Einsatz von KI in Form von maschinellem Lernen, insbesondere von Deep Convolutional Neural Networks zur Erkennung von Läsionen, konnten die Ergebnisse allerdings deutlich verbessert werden. Besondere Herausforderungen sind hierbei, dass sehr große Datensätze für das Trainieren der Netzwerkmodelle benötigt werden und oft eine Anpassung an lokale Gegebenheiten notwendig ist. Internationale Studien zeigen, dass durch den Einsatz von KI die Anzahl falsch-positiver Ergebnisse um 5,7 Prozent (USA) bzw. um 1,2 Prozent (Großbritannien) gesenkt werden konnte und bessere Ergebnisse erzielt wurden als bei einer Beurteilung durch medizinisches Fachpersonal allein.[188]

Einige kommerzielle Anbieter haben bereits Programme auf den Markt gebracht, wie beispielsweise Siemens Healthineers mit der interaktiven Entscheidungsunterstützung syngo.Breast Care bei der Mammografie. Die kommerziell angebotenen Programme, wie zum Beispiel die Software Transpara, bieten insbesondere eine Scoringfunktion, welche die Mammografiebilder in Bezug auf das Vorliegen verdächtiger Läsionen stratifiziert. Dies ermöglicht es dem radiologischen Fachpersonal, in kürzerer Zeit komplexe Datensätze zu analysieren, und unterstützt die Entscheidungsfindung mit Blick auf weitere Behandlungsschritte. Insbesondere bei unklaren Befunden bleibt das Einholen einer Zweitmeinung einer weiteren Expertin / eines weiteren Experten allerdings unerlässlich. Insgesamt befinden sich KI-gestützte Systeme auch für die Diagnostik von Brustkrebs noch in der Entwicklung.

Ein verantwortlicher Umgang mit solchen Diagnosetools erfordert eine sorgfältige Abwägung der verschiedenen Chancen und Risiken, die mit ihrem Einsatz sowohl für die behandelnden Ärztinnen und Ärzte selbst als auch für die Betroffenen verbunden sind.

Aus *ärztlicher* Sicht wie auch aus Betroffenenperspektive dürfte der größte Vorteil einer Nutzung dieser Instrumente darin bestehen, pathologische Veränderungen der Zell- und Gewebestrukturen früher als bisher erkennen zu können und damit die Möglichkeiten einer erfolgreichen, an die individuellen Gegebenheiten der jeweiligen Betroffenen angepassten Therapie zu verbessern. Der aus der früheren Detektion des

187 Sechopoulos/Teuwen/Mann 2021.
188 McKinney et al. 2020.

Tumors resultierende Zeitgewinn steigert nicht nur generell die Erfolgsaussichten der Behandlung, sie erweitert auch insofern die Therapieoptionen, als bestimmte Behandlungsmethoden (z. B. die operative Entfernung des Tumors) nur so lange möglich sind, wie eine Metastasierung noch nicht stattgefunden hat.

Ein weiterer Vorteil besteht darin, dass KI-gestützte Systeme die Auswertung digitaler Bilder beschleunigen und das ärztliche Personal von monotonen Routinearbeiten entlasten können, sodass diesem idealtypischerweise mehr Zeit für die Vermittlung der Befunde im persönlichen Gespräch zur Verfügung steht.[189]

Diesen Chancen stehen aber auch Risiken gegenüber. So können Ärztinnen und Ärzte infolge der fortschreitenden Delegation bestimmter Aufgaben an technische Systeme zum einen eigene Kompetenzen schleichend verlieren, da sie diese immer seltener selbst anwenden. Zum anderen könnten sie gerade aufgrund des verlorengegangenen eigenen Erfahrungswissens dazu neigen, ihre Sorgfaltspflichten im Umgang mit derartigen Instrumenten dadurch zu verletzen, dass sie deren Empfehlungen blind folgen (Automation Bias). Da auch KI-basierte Instrumente nicht fehlerfrei arbeiten, sondern mitunter nur andere Fehler als Menschen begehen und die Nichtentdeckung dieser Fehler fatale Konsequenzen haben kann, sollten KI-gestützte Entscheidungssysteme immer so gestaltet werden, dass die konkrete Art der Übermittlung ihrer Analyseergebnisse (etwa in Gestalt einer Wahrscheinlichkeitsangabe für das Vorliegen einer pathologischen Veränderung[190]) deutlich macht, dass hier noch eine ärztliche Plausibilitätsprüfung erforderlich ist. Da die Behandelnden moralisch und rechtlich

189 Baltzer 2021.

190 Ein Bespiel für dieses Problem ist ein Algorithmus, der in der Notfallaufnahme zur Detektion einer Hirnblutung eingesetzt wird und dessen Ausgabe lediglich ein „Ja“ oder „Nein“ umfasst. Hier liegt das Problem nicht in der Nutzung der KI als solcher, sondern in der undifferenzierten Art der Präsentation des Analyseergebnisses.

dafür verantwortlich sind, den Betroffenen die aus ihrer Sicht beste Behandlung anzubieten, gehört auch die kritische Überprüfung der Ergebnisse der von ihnen eingesetzten technischen Instrumente zu den ärztlichen Sorgfaltspflichten, die als solche nicht delegierbar ist.

Ein weiteres Beispiel aus der Praxis im Krankenhaus soll die angesprochene Problematik illustrieren.[191] Ein KI-Gerät in der Radiologie unterstützt das Erkennen von Hirnblutungen bei Patientinnen und Patienten, indem es eine binäre Klassifikation anzeigt (ja, Blutung liegt vor, bzw. nein, Blutung liegt nicht vor). Es stellt sich dabei die Frage, ob eine derart grobe Klassifikation ohne zusätzliche, differenziertere Angaben aus der Anwendungsperspektive eine sinnvolle Entscheidungsunterstützung darstellt. Sowohl zur Einschätzung der Sicherheit des binären Klassifikationsergebnisses durch das ärztliche Personal als auch mit Blick auf die weitere Behandlungsplanung wären komplexere Aussagen wünschenswert, etwa zur Wahrscheinlichkeit der (verschiedenen) Diagnosen, zur Lokalisation der Blutung im Gehirn oder zu möglichen Ursachen. Grundsätzlich erscheint ein Einsatz des Gerätes als maschineller „Zweitgutachter" aber durchaus sinnvoll.

Der Einsatz dieser KI-gestützten Klassifizierung stellt auch für erfahrene Fachkräfte trotz eines potenziellen Mehraufwandes eine Unterstützung dar, die zur besseren Urteilsbildung beiträgt. Praxisrelevant sind dabei technische Umsetzungsdetails wie beispielsweise eine direkte Integration im Scanner und eine schnelle Auswertung ohne gesonderte Einwilligung in den Datentransfer an anderer Stelle. Es sollte dabei sichergestellt werden, dass die KI-basierte Analyse das diagnostische Potenzial auf Grundlage der erhobenen Daten ausschöpft und den Behandelnden nachvollziehbar präsentiert und gegebenenfalls mit weiteren Daten (wie etwa typischen Vergleichsfällen und möglichen Konsequenzen der Läsion) in Verbindung bringt.

191 Kiefer/May 2022.

Für die *Patienten*perspektive ergeben sich noch weitere Aspekte. Denn obwohl die Chance einer verbesserten Früherkennung patientenseitig uneingeschränkt zu begrüßen ist, weil sie die Aussichten auf eine möglichst schonende Therapie und Heilung erhöht, bleibt die fortschreitende Technisierung einzelner Behandlungsschritte für die Arzt-Patienten-Beziehung keineswegs folgenlos. Personen, die im Rahmen ihrer medizinischen Behandlung mit neuen Diagnostikinstrumenten konfrontiert werden, haben erfahrungsgemäß Fragen, etwa zur Sicherheit und Genauigkeit oder zum Datenschutz, und Befürchtungen, beispielsweise nicht mehr als Person wahrgenommen, sondern auf ihre Daten reduziert zu werden. Diese Fragen und Befürchtungen sind ernst zu nehmen und in einer umfassenden Aufklärung gezielt zu thematisieren, um den Patientinnen und Patienten auch die Möglichkeit zur Einholung einer Zweitmeinung zu eröffnen. Dieser Aufklärungsbedarf steigt zwangsläufig an, wenn nicht nur ein einzelnes Segment der ärztlichen Handlungssequenz technisch ersetzt wird, sondern weitere Ersetzungen möglich werden.

Da das Vertrauen in die Behandlung patientenseitig maßgeblich von der personalen Zuwendung der Behandelnden abhängt, dürfte nicht davon auszugehen sein, dass diese mit der fortschreitenden Implementierung von KI-Systemen in verschiedene Schritte des komplexen Behandlungsprozesses überflüssig werden. Vielmehr besteht teilweise die Hoffnung, die schon heute bestehende Tendenz der Anonymisierung partiell zu korrigieren und den Ärztinnen und Ärzten wieder mehr Freiräume zur Erbringung ihrer eigentlichen Aufgaben zu schaffen.[192]

Ersetzungen mittleren Ausmaßes

Da sich die digitale Transformation der Medizin dynamisch entwickelt/vollzieht, gibt es bereits heute eine Reihe von Phänomenen, bei denen die Grenze zwischen einer engen und

192 Health Education England 2019.

einer mittleren Ersetzung einzelner Handlungsschritte durch KI-Komponenten zunehmend verschwimmt und die darauf hindeuten, dass es sich hierbei insgesamt um ein Kontinuum unterschiedlicher Nutzungsmöglichkeiten solcher Tools handelt, die zum gegenwärtigen Zeitpunkt erst teilweise realisiert sind.

Ein Beispiel hierfür ist ein Algorithmus zur Unterstützung des anästhesiologischen Fachpersonals zur Berechnung des Risikos, während einer Operation zu versterben. In einem ersten Schritt kann man die Nutzung solcher Instrumente als eine begrüßenswerte Erweiterung des ärztlichen Fähigkeitenspektrums deuten, da das bisherige operationsbegleitende Monitoring der Vitalparameter der operierten Person durch die Fachleute hier technisch so substituiert wird, dass eine frühzeitige Prognose kritischer Zustände möglich ist und eine entsprechende ärztliche Reaktion eingeleitet werden kann. Es ist aber in einem zweiten Schritt auch denkbar, dass der Algorithmus künftig nicht nur das Sterblichkeitsrisiko der behandelten Person in Echtzeit berechnet, sondern zusätzlich auch noch automatisiert die jeweils situativ angezeigten, bislang dem anästhesiologischen Personal vorbehaltenen Veränderungen der Operationsbedingungen (z. B. die Dosierung des Narkosemittels oder die zusätzliche Gabe von Katecholaminen) auslösen könnte.

In der Folge könnte sich das ärztliche Aufgabenspektrum künftig zunehmend auf die Wahrnehmung prä- und postoperativer Aufgaben (z. B. die Aufklärung der Patientinnen und Patienten) konzentrieren. Unabhängig davon, wie realistisch dieses Szenario ist, zeigen solche Beispiele, dass es eine Reihe von Misch- und Übergangsphänomenen gibt, die unter anderem auch daraus resultieren, dass die verschiedenen funktional differenzierbaren Segmente ärztlichen Handelns Teile eines einheitlichen Handlungszusammenhangs bilden.

Doch ist es nicht nur die innere Einheit von Diagnostik und Therapie, die solche Übergänge begünstigt. Auch externe Rahmenbedingungen und Effizienzüberlegungen zur

Optimierung der Prozessqualität komplexer Behandlungsabläufe können Treiber einer fortscheitenden Ersetzungsdynamik sein (siehe Infokasten 3).

Infokasten 3: Mittlere Ersetzung am Beispiel von Prostatakrebs

Prostatakrebs (Prostatakarzinom) ist mit rund 68600 Neuerkrankungen jährlich die häufigste Krebserkrankung beim Mann.[193] Die Ursache dieser Krebserkrankung ist noch weitgehend unbekannt, allerdings ist einer der wichtigsten Risikofaktoren das Alter. Aufgrund des demografischen Wandels hat die Anzahl an Prostatakrebserkrankungen deutlich zugenommen; gleichzeitig haben sich die Überlebensaussichten aber erheblich verbessert, da etwa zwei Drittel der Fälle bereits im Anfangsstadium diagnostiziert werden. In Screeningverfahren zur Früherkennung wird das von den Karzinomzellen verstärkt gebildete prostataspezifische Antigen (PSA) gemessen.

Erhebliche Fortschritte in der Prostatakrebsdiagnostik wurden zudem im Bereich der multiparametrischen Magnetresonanztomografie (mpMRT) in Kombination mit ultraschallgesteuerter Biopsie erzielt. Der KI-Einsatz unterstützt die Fusion der umfangreichen Datensätze und verbessert mittels Anwendung von Deep Convolutional Neural Networks als Faltungsnetzwerk zur Bildsegmentierung, sogenannten U-Nets, die Erkennung und Lokalisation von suspekten Läsionen. Automatisierte Bildfusionsverfahren unterstützen die Gewebeentnahme, indem zuvor erstellte mpMRT-Aufnahmen in Echtzeit mit Ultraschallaufnahmen bei der Gewebeentnahme überlagert werden. Mit dieser Methode konnte durch eine präzise Biopsie die Entdeckung behandlungsbedürftiger Prostatakarzinome von ca. 50 Prozent auf über 90 Prozent erhöht werden.[194] Für eine Therapieentscheidung ist letztlich jedoch eine pathologische Diagnose unerlässlich, die sich bisher weitgehend auf mikroskopische Untersuchungen stützt. Zunehmend werden aber, ähnlich der Auswertung von bildgebenden Verfahren wie Magnetresonanztomografie und Ultraschall, auch mikroskopische Bilder von Gewebeschnitten digitalisiert und automatisiert mit Unterstützung von KI-Algorithmen ausgewertet. Insbesondere bei der integrativen Auswertung aller Befunde kann KI darüber hinaus helfen, richtige Entscheidungen zu treffen, indem zunächst anhand eines „digitalen Zwillings" Therapieoptionen getestet werden, um möglichst erfolgversprechende Ansätze auszuwählen.

In der bildgeführten Therapie des Prostatakarzinoms kommen weitere (teil-) automatisierte Verfahren der Chirurgie und Strahlentherapie hinzu. Inzwischen werden in Deutschland an über 140 Kliniken Operationen mit dem Da-Vinci-Operationssystem durchgeführt.[195] Das roboterassistierte Operieren erweitert das Spektrum minimalinvasiver Operationen im Bauchraum. Die

193 https://www.krebsdaten.de/Krebs/DE/Content/Krebsarten/Prostatakrebs/prostatakrebs_node.html [03.03.2023].

194 Ahmed et al. 2017; Schelb et al. 2019. Vgl. Schlemmer/Hohenfellner 2021; Gigerenzer/Mata/Frank 2009.

195 https://klinikradar.de/da-vinci-op-roboter/kliniken [11.01.2023]; https://www.klinikkompass.com/kliniken-mit-einem-da-vinci-roboter [11.01.2023].

Geräte werden vom ärztlichen Personal gesteuert, das am Monitor die Gefäß- und Gewebestrukturen deutlich besser erkennen kann als mit bloßem Auge; die Schnitte werden von Roboterarmen ausgeführt. So wird die Präzision bei minimalinvasiven Operationen deutlich erhöht. Dies ermöglicht eine radikale und doch schonende und nervenerhaltende Prostataoperation (Prostatektomie).[196] Eine Therapiealternative bei Prostataoperationen, die sich noch in der Erprobungsphase befindet, ist die Cyberknife-Methode, eine vergleichsweise neue radiochirurgische Methode zur Bestrahlung von kleinen, gut lokalisierten Tumoren. Dabei werden Röntgenstrahlen gebündelt und höchst präzise auf den Tumor gelenkt, sodass das Strahlenbündel mit enormer Kraft die Funktion eines Operationsmessers übernimmt. Während die operierende Person die Strahlung reguliert, wird die Position in Echtzeit angepasst, was eine präzise Behandlung ermöglicht.[197]

Eine ethische Reflexion hat zunächst zu berücksichtigen, dass die angemessene Versorgung einer steigenden Anzahl von Prostatakrebspatienten vor einer doppelten strukturellen Herausforderung steht. Einerseits müssen die Arbeitsabläufe von Diagnostik und Therapie mit dem Ziel einer besseren Vernetzung aller beteiligten Akteure auf den unterschiedlichen Versorgungsebenen und eines beschleunigten Datenaustauschs weiter standardisiert werden. Andererseits soll die Behandlung für die einzelne Person durch eine bessere Stratifizierung relevanter Subgruppen personalisiert werden, um nutzlose oder sogar schädliche Therapieversuche zu vermeiden.

Nach Einschätzung vieler Sachverständiger[198] kann die verstärkte Nutzung von KI-Instrumenten die notwendigen Entscheidungen während des gesamten Behandlungsprozesses positiv unterstützen, da sich die einzelnen Segmente – von der klinischen Untersuchung und Labordiagnostik über Bildgebung und Biopsie bis hin zu Therapie und Nachsorge – in Gestalt von Entscheidungsbäumen strukturieren und damit prinzipiell einer Bearbeitung durch KI-Instrumente zugänglich machen lassen. Dabei sind es keineswegs nur reine

196 https://www.leading-medicine-guide.de/behandlung/da-vinci-prostatektomie [11.01.2023].
197 https://www.leading-medicine-guide.com/de/behandlung/cyberknife [11.01.2023].
198 Zentrale Ethikkommission bei der Bundesärztekammer 2021, A2 ff.

Effizienzüberlegungen, die für den verstärkten Einsatz von KI-Systemen in diesem Bereich sprechen, sondern vor allem die dadurch ermöglichte Verbesserung der Behandlungsqualität, von der letztlich alle Beteiligten profitieren.

Mit den empirisch nachweisbaren Vorteilen – sowohl bei der Früherkennung von Prostatakrebs als auch bei der Therapie durch neue KI-gesteuerte Operationsroboter, die mittlerweile eigenständig nicht nur die Ränder von Tumorgewebe präziser als ein Mensch identifizieren, sondern auch besonders gewebeschonend innerhalb des Organismus manövrieren können – gehen aber auch verschiedene Herausforderungen einher, die für das hier vorgestellte Beispiel ebenso gelten wie für viele ähnliche Anwendungsszenarien von KI im Gesundheitsbereich, bei denen menschliche Tätigkeiten in engem bis mittlerem Ausmaß durch KI-gestützte Technik ersetzt werden.

Erstens ist für eine verlässliche Unterstützung der Diagnose und Therapie durch KI eine flächendeckende und einheitliche – oder zumindest qualitativ vergleichbare – technische Ausrüstung sowie entsprechend geschultes Personal und eine kontinuierliche Qualitätssicherung unabdingbar.

Zweitens ist aufgrund der immanenten Grenzen von KI-Systemen davon auszugehen, dass Ärztinnen und Ärzte bei der Diagnose und Behandlung von Prostatakarzinomen auch langfristig unverzichtbare Funktionen zu erfüllen haben, die nicht maschinell substituierbar sind. Beispielsweise müssen Befunde auf Plausibilität geprüft werden.

Drittens bedarf die konkrete Diagnose- und Therapieplanung einer Berücksichtigung der umfassenden Lebenssituation jedes einzelnen Betroffenen, seiner jeweiligen Vorerkrankungen, individuellen Interessen und Wertpräferenzen, zu der auch fortschrittliche KI-Systeme nicht in der Lage sind.

Viertens scheint es ärztlicherseits erforderlich, eine alle Versorgungsebenen und Behandlungsschritte einbeziehende Kommunikationsstrategie zu entwickeln, um der Gefahr zu wehren, dass die jeweiligen Behandelnden immer mehr hinter der Technik verschwinden und sich die von ihnen behandelten

Personen mit ihren Fragen und Ängsten zunehmend allein gelassen fühlen. Da über den Einsatz solcher Systeme im Vorfeld aufgeklärt werden muss und auch während der Nutzung auftauchende Fragen zum Beispiel zur verantwortlichen Datennutzung, zum Datenschutz oder zur Zuverlässigkeit der maschinell produzierten Empfehlungen beantwortet werden müssen, dürfte damit zu rechnen sein, dass die kommunikativen Anforderungen an eine vertrauensvolle Arzt-Patienten-Beziehung in der Zukunft sogar noch weiter zunehmen.

Fünftens ist damit zu rechnen, dass sich bestimmte Personen dem Einsatz dieser technischen Systeme selbst dann verweigern werden, wenn sie bisherigen Behandlungsformen aus medizinischer Sicht überlegen sind, sodass es im Rahmen der Interaktion zwischen ärztlichem Personal und den ihnen anvertrauten Personen klarer Regeln bedarf, die nicht nur die Patientenautonomie, sondern auch die ärztliche Therapiefreiheit schützen.

Sechstens ist zu prüfen, wie sich die immer stärkere Implementierung von KI-Systemen auf die Kostenentwicklung im Gesundheitswesen auswirkt.

Siebtens sind Herausforderungen zu beachten, die sich bei der für die (Weiter-)Entwicklung und Evaluierung solcher Systeme notwendigen Sammlung, Verarbeitung und Weitergabe von gesundheitsbezogenen Daten ergeben. Im Bereich der KI-Nutzung ist das vom Deutschen Ethikrat schon 2017 mit Blick auf die gesundheitsbezogene Datennutzung formulierte, übergreifende Ziel der Datensouveränität besonders schwierig zu erreichen.[199] Zum einen erfordern nahezu alle KI-Nutzungen im Gesundheitsbereich enorm große Big-Data-Datensätze. Die mitunter sehr restriktive individuelle Auslegungspraxis geltender Datenschutzbestimmungen und einzelne teils überholte

199 Deutscher Ethikrat 2017, 251–280. Datensouveränität wird hier verstanden „als eine den Chancen und Risiken von Big Data angemessene verantwortliche informationelle Freiheitsgestaltung“ (ebd., 252).

Datenschutzregeln[200] können deren Gewinnung und damit auch sinnvollen Entwicklungen entgegenstehen. Zum anderen erlauben die Anwendungen oft noch unmittelbareren Zugriff auf die Privatsphäre von Betroffenen als andere Anwendungen der datenreichen Medizin. Darüber hinaus besteht das Problem der Verwendung solcher Daten für andere Zwecke. Entsprechend entbindet das Ziel, neue Erkenntnisse für bessere zukünftige Behandlungsmethoden im Bereich der öffentlichen Forschung zu generieren, keinesfalls davon, die Personen, die ihre Daten für die Forschung zur Verfügung stellen, möglichst umfassend (gegebenenfalls auch über den Grad der Fremdnützigkeit des jeweiligen Forschungsprojektes) aufzuklären. Erschwerend kommt hinzu, dass etwa die Sekundärnutzung von klinischen Daten für KI-getriebene klinische Forschung in einem komplexen Forschungsfeld erfolgt, in dem die Grenzen zwischen der im Gesundheitswesen verbreiteten, rein öffentlichen oder kooperativen öffentlich-privaten Forschung mit Gemeinwohlorientierung einerseits und der Forschung von Akteuren ganz außerhalb dieses insgesamt stark regulierten Systems andererseits zunehmend verschwimmen.

Weitreichende Ersetzung

Einer der wenigen medizinischen Handlungsbereiche, in denen KI-basierte Systeme zum Teil ärztliches bzw. anderes Gesundheitspersonal jedenfalls in bestimmten Kontexten und Gruppen bereits weitgehend oder sogar vollständig ersetzen, ist die Psychotherapie. In den letzten Jahren ist eine Fülle von Instrumenten zur (Teil-)Diagnose und Behandlung verschiedener psychischer Probleme entstanden, meist in Form von frei erhältlichen bildschirmbasierten Apps, etwa Chatbots, mit denen auf algorithmischer Basis eine Art von Therapie

200 Ein Beispiel hierfür ist Art. 27 Abs. 4 Satz 6 des Bayerischen Krankenhausgesetzes, der erst im Juni 2022 aufgehoben wurde und es bis dahin bayerischen Kliniken nicht erlaubte, dass Patientendaten das Krankenhaus verlassen.

– zumeist kognitive Verhaltenstherapie – mit den Betroffenen abläuft.[201]

Vielfach werden solche Instrumente als Elemente einer ärztlich bzw. therapeutisch geführten Behandlung eingesetzt, etwa um durch ein kontinuierliches Monitoring Klarheit über die (auch kontextspezifische) Ausprägung, Variabilität und Häufigkeit von Symptomen zu gewinnen (sogenanntes Ecological Momentary Assessment als Alternative zu retrospektiven Selbstberichten), Aspekte bestimmter Therapien zu verstärken (z. B. Biofeedback oder Virtual Reality als imaginäre Konfrontationstherapie), Verhaltensweisen zu üben oder therapeutisch wirksame Aufgaben zu erledigen (z. B. Entspannungsübungen oder Selbstinstruktionen).[202] In dieser Form bleiben diese Anwendungen also unter einer gewissen professionellen Aufsicht und unterscheiden sich wenig von den bereits seit einigen Jahrzehnten bekannten therapeutischen Begleitinstrumenten, wie etwa papier- oder bildschirmbasierten psychotherapeutischen Aufgaben auf Fragebogen- oder Entscheidungsbaumbasis. Hier werden also allenfalls einzelne Elemente im Sinne einer engen Ersetzung auf algorithmischer Basis übernommen, gegebenenfalls mit etwas mehr Benutzungsfreundlichkeit und einer Menge (ihrerseits ethisch nicht immer unproblematischen) Möglichkeiten der Datensammlung[203], ohne dass die therapeutische Beziehung notwendigerweise grundlegend verändert würde.

Allerdings sind viele dieser Apps frei in App-Stores erhältlich und werden daher auch außerhalb des medizinisch-therapeutischen Kontextes eingesetzt. In dieser Form erfolgt – wenn Betroffene solche Apps nicht nur als einen niedrigschwelligen Einstieg in eine therapeutische Behandlung durch

201 Übersichten in Fiske/Henningsen/Buyx 2019; Gratzer/Goldbloom 2020.
202 Van Daele et al. 2021.
203 Bei verschiedenen in App-Stores erhältlichen Apps ist unklar, ob und gegebenenfalls wie Nutzungsdaten weiterverwendet werden. Da Anwendungen wie etwa der Chatbot Woebot kostenlos sind, ist anzunehmen, dass Daten weiterverkauft oder jedenfalls einer kommerziellen Nutzung zugeführt werden.

medizinisches Fachpersonal nutzen – tatsächlich eine Therapie, ganz ohne dass menschliches Personal hinzukäme.[204] Die Chatbots treten dabei in „Gesprächskontakt" und „diagnostizieren" den psychischen Zustand, in dem sich eine Person befindet. Sie bieten verschiedene therapeutische Wege sowie Übungen an und interagieren regelmäßig und aufsuchend mit den sie nutzenden Personen (es sei denn, diese haben die entsprechende Funktion ausgestellt).[205] Die meisten dieser Chatbots weisen darauf hin, dass etwa suizidale Gedanken unmittelbare, intensivere Behandlung erfordern, und „schlagen vor", Krisentelefon-Services in Anspruch zu nehmen. Sie sind jedoch nicht mit Notdiensten oder Ähnlichem verbunden und fungieren daher nicht als eigenständiges Warnsystem.

Hier ergibt sich also tatsächlich eine weite Ersetzung insofern, als die Chatbots und diejenigen, die mit ihnen interagieren, in einer Art direkten, therapeutischen Beziehung stehen. Bisher gibt es wenige empirische Studien oder theoretische Abhandlungen, die dieses Phänomen untersuchen.[206] Dennoch liegen ethische Vorteile und Bedenken auf der Hand. Positiv diskutiert wird insbesondere, dass solche Apps angesichts ihrer Niedrigschwelligkeit und ständigen Verfügbarkeit Menschen in Erstkontakt mit therapeutischen Angeboten bringen können, die sonst oft zu spät oder gar keine Therapie erhalten. Dies gilt insbesondere für Gruppen, die oft schwer (z. B. wohnortbedingt) mit anderen Angeboten erreicht werden (Gateway-Phänomen).[207]

Hinzu kommt ein weithin für die Nutzung etwa von Aufklärungsalgorithmen in Krankenhäusern beschriebenes Phänomen[208], dass Menschen maschinellen „Therapeuten" mehr von sich berichten und ihnen gegenüber weniger Scham verspüren als im Gespräch von Mensch zu Mensch. Gerade

204 Ebert et al. 2018.
205 Mehta et al. 2021.
206 Holohan/Fiske 2021.
207 Parviainen/Rantala 2022.
208 Pugh 2018; Dennis et al. 2020.

Scham oder die Sorge vor Stigmatisierung hält bekanntermaßen viele Menschen davon ab, frühzeitig Therapieangebote aufzusuchen. Diesbezüglich könnte also gerade die Ersetzung menschlichen Fachpersonals dazu führen, dass überhaupt ein therapieähnliches Angebot in Anspruch genommen wird. Hier ist allerdings zu berücksichtigen, dass die Frage, ob und wieweit sich Menschen öffnen – und welche Effekte dies hat –, nicht nur von der Anonymität der Interaktion, sondern auch vom Vertrauen in die Kontaktperson abhängt. Der anonyme Chatbot ersetzt nicht ohne Weiteres das als vertrauenswürdig wahrgenommene therapeutische Fachpersonal, denn die Wirkung von Psychotherapie entfaltet sich in vielen Fällen erst auf der Grundlage der Therapeut-Klienten-Beziehung.[209] Und schließlich wird insbesondere mit Blick auf Regionen, in denen nicht ausreichend Therapieplätze zur Verfügung stehen, argumentiert, dass angesichts des immer weiter steigenden Versorgungsbedarfs[210] gerade milder psychischer Probleme Chatbot-Apps jedenfalls „besser als nichts" sind.[211]

Allerdings ergeben sich insbesondere im Falle einer Ersetzung therapeutischer Fachkräfte durch Maschinen aus ethischer Sicht wichtige Bedenken und Probleme. Offenkundig sind Probleme wie die mangelnde Qualitätskontrolle der Bots (vor allem wenn sie nicht als digitale Gesundheitsanwendungen[212] zugelassen sind), Fragen zur Datensammlung und -weiterverwendung, Fragen zum Schutz der Privatsphäre sowie die erwähnte fehlende Warnfunktion etwa bei eindeutiger Suizidalität.

Hinsichtlich der Betroffenenperspektive kann zudem angenommen werden, dass gerade vulnerable Personen, die ohnehin schlecht versorgt sind, den Eindruck gewinnen, mit vermeintlich zweitklassigen therapeutischen Surrogaten

209 Croes/Antheunis 2021; Gerhardinger 2020, 225 ff.
210 Vgl. auch Deutscher Ethikrat 2022a.
211 Fiske/Henningsen/Buyx 2019, 5.
212 https://www.bfarm.de/DE/Medizinprodukte/Aufgaben/DiGA-und-DiPA/DiGA/_node.html [22.02.2023].

abgespeist zu werden. Weiterhin kann nicht unterstellt werden, dass alle, die solche Apps verwenden, umfassend über deren Charakter informiert sind. Es ist möglich, dass manche die App mit einer Art telemedizinischem Angebot verwechseln und davon ausgehen, dass auf der anderen Seite doch ein Mensch therapeutisch agiert. Hier würde also die Ersetzung als solche verkannt – mit potenziell problematischen Folgen.

Schließlich ist noch so gut wie gar nicht untersucht, ob innerhalb der therapeutischen Beziehung zwischen einer App und denjenigen, die sie nutzen, Phänomena von Übertragung – ein wesentliches Element bei vielen therapeutischen, insbesondere psychoanalytischen Ansätzen – erfolgen. Es ist zumindest anzunehmen, dass Menschen eine Art emotionale Beziehung zur therapeutischen App aufbauen könnten, die von Zuneigung, Sich-Verlassen bis hin zu Abhängigkeitsaspekten reichen könnte. Daraus ergibt sich nicht nur die Frage, ob und inwieweit hier Empathie und Verstehen technisch substituiert werden können oder sollen, sondern auch, ob dies auf Dauer Effekte auf die Beziehungsfähigkeiten der Betroffenen haben könnte. Therapiebegünstigende (weil reflektierte) Gegenübertragungen seitens der therapeutischen Fachkraft können von KI-Anwendungen hingegen nicht erwartet werden. Dies verdeutlicht, dass digitale Anwendungen auf dem Gebiet der Psychotherapie nicht unabhängig von der jeweiligen Therapieform betrachtet werden können – ebenso wenig, wie die Therapieformen unabhängig von der Persönlichkeit derjenigen, die sie nutzen, und der jeweils infrage stehenden Problematik sowie ihrem unmittelbaren Bedrohungspotenzial betrachtet werden können.

Auch aus gesellschaftlicher Perspektive ergeben sich Fragen. Kontrovers diskutiert wird etwa, ob die zunehmende Nutzung solcher Apps einem weiteren Abbau von therapeutischem Fachpersonal Vorschub leistet und damit die Reduktion von Versorgungsbereichen beschleunigt. Ebenso werden eine Verstärkung von gesundheitlichen Ungleichheiten und *digital divides* sowie auch ein möglicher weiterer Prestigeverlust der

sprechenden Medizin, die in ungerechtfertigter Weise als durch Algorithmen einfach ersetzbar erscheinen könnte, angemahnt.

5.3 Fazit und Empfehlungen

Die Erfahrung, dass sich die digitale Transformation unseres Gesundheitssystems auf verschiedene Bereiche des medizinischen Handelns bislang sehr unterschiedlich auswirkt, bestätigt sich auch im Blick auf den speziellen Bereich von KI-Anwendungen. Dies gilt nicht allein für den Grad der Durchdringung eines Handlungsfeldes mit den neuen technischen Möglichkeiten und die Dynamik der daraus resultierenden Veränderungen für die Arzt-Patienten-Beziehung, sondern auch für die Gründe, die diese Entwicklung jeweils vorantreiben.

Neben naheliegenden immanenten Faktoren wie der Existenz immer größerer Datenmengen, die einer maschinellen Bearbeitung zugänglich sind, können auch kontingente Faktoren die Nutzung von KI-Anwendungen begünstigen. Dazu gehören beispielsweise Versorgungsengpässe aufgrund von Personalmangel, aber auch die Möglichkeit bzw. Erwartung präziserer Diagnostik oder neuerer, gegebenenfalls niedrigschwelligerer therapeutischer Anwendungen in einzelnen Bereichen.

Die in den vorigen Abschnitten vorgestellten Beispiele machen deutlich, dass es dabei sehr auf die jeweils konkreten Bedingungen des Einsatzes von KI-basierten Tools ankommt, ob diese beispielsweise innerhalb oder außerhalb einer etablierten Arzt-Patienten-Beziehungen zum Einsatz kommen. Bei der Zertifizierung zukünftiger Anwendungen sind solche kontextspezifischen Fragen zu berücksichtigen.

Ungeachtet der daraus resultierenden Notwendigkeit einer Binnendifferenzierung des weiten Handlungsfeldes der Medizin liegen einige übergreifenden Empfehlungen für den Gesundheitssektor nahe.

Empfehlungen

- *Empfehlung Medizin 1:* Bei der Entwicklung, Erprobung und Zertifizierung medizinischer KI-Produkte bedarf es einer engen Zusammenarbeit mit den relevanten Zulassungsbehörden sowie insbesondere mit den jeweils zuständigen medizinischen Fachgesellschaften, um Schwachstellen der Produkte frühzeitig zu entdecken und hohe Qualitätsstandards zu etablieren.
- *Empfehlung Medizin 2:* Bei der Auswahl der Trainings-, Validierungs- und Testdatensätze sollte über bestehende Rechtsvorgaben hinaus mit einem entsprechenden Monitoring sowie präzise und zugleich sinnvoll umsetzbaren Dokumentationspflichten sichergestellt werden, dass die für die betreffenden Patientengruppen relevanten Faktoren (z. B. Alter, Geschlecht, ethnische Einflussfaktoren, Vorerkrankungen und Komorbiditäten) hinreichend berücksichtigt werden.
- *Empfehlung Medizin 3:* Bei der Gestaltung des Designs von KI-Produkten zur Entscheidungsunterstützung ist sicherzustellen, dass die Ergebnisdarstellung in einer Form geschieht, die Gefahren etwa von Automatismen (Automation Bias) transparent macht, ihnen entgegenwirkt und die die Notwendigkeit einer reflexiven Plausibilitätsprüfung der jeweils vom KI-System vorgeschlagenen Handlungsweise unterstreicht.
- *Empfehlung Medizin 4:* Bei der Sammlung, Verarbeitung und Weitergabe von gesundheitsbezogenen Daten sind generell strenge Anforderungen und hohe Standards in Bezug auf Aufklärung, Datenschutz und Schutz der Privatheit zu beachten. In diesem Zusammenhang verweist der Deutsche Ethikrat auf seine 2017 im Kontext von Big Data und Gesundheit formulierten Empfehlungen, die sich am Konzept der Datensouveränität orientieren, das für den Bereich von KI-Anwendungen im Gesundheitsbereich gleichermaßen Gültigkeit entfaltet.

» *Empfehlung Medizin 5:* Bei durch empirische Studien sorgfältig belegter Überlegenheit von KI-Anwendungen gegenüber herkömmlichen Behandlungsmethoden ist sicherzustellen, dass diese allen einschlägigen Patientengruppen zur Verfügung stehen.

» *Empfehlung Medizin 6:* Für erwiesen überlegene KI-Anwendungen sollte eine rasche Integration in die klinische Ausbildung des ärztlichen Fachpersonals erfolgen, um eine breitere Nutzung vorzubereiten und verantwortlich so gestalten zu können, dass möglichst alle Patientinnen und Patienten davon profitieren und bestehende Zugangsbarrieren zu den neuen Behandlungsformen abgebaut werden. Dazu ist die Entwicklung einschlägiger Curricula/Module in Aus-, Fort- und Weiterbildung notwendig. Auch die anderen Gesundheitsberufe sollten entsprechende Elemente in die Ausbildung aufnehmen, um die Anwendungskompetenz bei KI-Anwendungen im Gesundheitsbereich zu stärken.

» *Empfehlung Medizin 7:* Bei routinemäßiger Anwendung von KI-Komponenten sollte nicht nur gewährleistet werden, dass bei denjenigen, die sie klinisch nutzen, eine hohe methodische Expertise zur Einordnung der Ergebnisse vorhanden ist, sondern auch strenge Sorgfaltspflichten bei der Datenerhebung und -weitergabe sowie bei der Plausibilitätsprüfung der maschinell gegebenen Handlungsempfehlungen eingehalten werden. Besondere Aufmerksamkeit erfordert die Gefahr eines Verlustes von theoretischem wie haptisch-praktischem Erfahrungswissen und entsprechenden Fähigkeiten (Deskilling); dieser Gefahr sollte mit geeigneten, spezifischen Fortbildungsmaßnahmen entgegengewirkt werden.

» *Empfehlung Medizin 8:* Bei fortschreitender Ersetzung ärztlicher, therapeutischer und pflegerischer Handlungssegmente durch KI-Komponenten ist nicht nur sicherzustellen, dass Patientinnen und Patienten über alle entscheidungsrelevanten Umstände ihrer Behandlung vorab

informiert werden. Darüber hinaus sollten auch gezielte kommunikative Maßnahmen ergriffen werden, um dem drohenden Gefühl einer zunehmenden Verobjektivierung aktiv entgegenzuwirken und das Vertrauensverhältnis zwischen den beteiligten Personen zu schützen. Je höher der Grad der technischen Substitution menschlicher Handlungen durch KI-Komponenten ist, desto stärker wächst der Aufklärungs- und Begleitungsbedarf der Patientinnen und Patienten. Die verstärkte Nutzung von KI-Komponenten in der Versorgung darf nicht zu einer weiteren Abwertung der sprechenden Medizin oder einem Abbau von Personal führen.

» *Empfehlung Medizin 9:* Eine vollständige Ersetzung der ärztlichen Fachkraft durch ein KI-System gefährdet das Patientenwohl und ist auch nicht dadurch zu rechtfertigen, dass schon heute in bestimmten Versorgungsbereichen ein akuter Personalmangel besteht. Gerade in komplexen Behandlungssituationen bedarf es eines personalen Gegenübers, das durch technische Komponenten zwar immer stärker unterstützt werden kann, dadurch selbst als Verantwortungsträger für die Planung, Durchführung und Überwachung des Behandlungsprozesses aber nicht überflüssig wird.

6 BILDUNG

6.1 Einleitung

Auch im Bildungsbereich kommen zunehmend digitale Technologien und algorithmische Systeme zum Einsatz.[213] Dies kann sowohl zur Standardisierung von Lernprozessen führen als auch im Falle von stark datenbasierten KI-Systemen – ähnlich wie in der Medizin – mehr Personalisierung ermöglichen, das heißt einen stärkeren Zuschnitt auf individuelle Bedingungen und Neigungen der Lernenden. Personalisiertes Lernen und Lehren werden dabei unter anderem dadurch möglich, dass durch umfangreiche maschinelle Datenerfassung und -auswertung ein Vielfaches an Informationen zum Lerngeschehen im Vergleich zu früheren Möglichkeiten zur Verfügung steht und von Lehrkräften, Lernenden und weiteren Beteiligten zur Personalisierung von Lehr- und Lernprozessen verwendet werden kann.[214] Wie auch im Anwendungsfeld Medizin reichen die Einsatzmöglichkeiten von sehr eng umrissenen punktuellen Angeboten für Lernende und Lehrkräfte bis hin zu Szenarien, in denen digitale Werkzeuge umfassend eingesetzt werden und beispielsweise datenbasierte, KI-gestützte Lehr- und Lernsysteme zeitweise oder gänzlich eine Lehrkraft ersetzen können.[215] Die damit verbundenen Veränderungspotenziale, Chancen und Risiken werden hier für die Schule als

213 OECD 2021.

214 Dies ist natürlich nur unter der Voraussetzung gegeben, dass sowohl mobile Endgeräte als auch kostenloses WLAN für die Schülerinnen und Schüler von staatlicher Seite zur Verfügung gestellt werden. Bildungsungleichheit bei den Startchancen darf dabei nicht weiter vertieft, sondern könnte durch das spezifische Eingehen auf die Heterogenität der Schülerschaft möglicherweise grundlegend angegangen werden.

215 https://www.egovernment-computing.de/ki-als-helfende-lehrkraft-waehrend-des-lockdowns-a-926866 [11.01.2023].

exemplarischem Lernort und Ort der Persönlichkeitsbildung vorgestellt.[216]

Jenseits der auch anderweitig thematisierten technischen Herausforderungen rund um die Nutzung digitaler und algorithmischer Systeme in der Schule[217] gilt das Interesse in dieser Stellungnahme vor allem den anthropologischen und ethischen Fragen. Im Mittelpunkt steht hierbei die Auseinandersetzung damit, wie die maschinelle Ersetzung bestimmter menschlicher Handlungssegmente Lehr- und Lernprozesse verändert, Handlungsmöglichkeiten aller Akteure erweitert oder auch vermindert. Zudem ist die Frage bedeutsam, wie dies das Selbstverständnis und Miteinander der verschiedenen Beteiligten am Lernort Schule und dem Zuhause der Schulpflichtigen beeinflusst und gegebenenfalls darüber hinaus gesellschaftliche Auswirkungen entfaltet.

Wie in Kapitel 3 ausgeführt, orientiert sich das hier zugrunde gelegte Verständnis von Bildung an der Fähigkeit des Menschen zu freiem und vernünftigem Handeln, das nicht auf behavioristische oder funktionalistische Modelle zu reduzieren ist. Die vielfältigen Möglichkeiten des Einsatzes digitaler und algorithmischer Angebote in der Bildung sind auf diese grundlegende Ausrichtung von Bildung zu beziehen. Wenn daher von Chancen und Risiken datenbasierter, KI-gestützter Lehr- und Lernsysteme die Rede ist, sind diese nicht einfach an technologischen Optimierungsvorstellungen zu orientieren. Alle technologischen Möglichkeiten der Gestaltung der Bildungsprozesse sind vielmehr daraufhin zu überprüfen, ob sie dem eingangs entwickelten Verständnis des Menschen als einer zur Selbstbestimmung und Verantwortung fähigen Person entsprechen oder ob sie diesem Verständnis entgegenstehen. Die Notwendigkeit einer solch grundlegende Betrachtung wird auch in den aktuellen Debatten rund um ChatGPT

216 Somit sind andere Orte der Bildung und des Lernens wie Hochschulen, Kindertagesstätten oder andere Formen der Aus- und Weiterbildung nicht explizit berücksichtigt.

217 Vgl. zum Überblick OECD 2021.

sichtbar.[218] Unmittelbar nachdem dieser KI-basierte und hoch performante Textgenerator frei zugänglich gemacht wurde, wurden disruptive Entwicklungen nicht zuletzt für den Bildungsbereich diskutiert. Ging es vordergründig zunächst um die Frage, wie Prüfungen fair gestaltet werden können, wenn man nicht weiß, ob Schülerinnen und Schüler ihre Hausarbeiten selbst geschrieben haben oder von ChatGPT haben produzieren lassen, so erfordert ChatGPT darüber hinaus eine erneute Auseinandersetzung damit, was Bildung ist und sein soll. Es geht um eine Vergewisserung und Neubestimmung dessen, was Ziel und Wert von Bildung sind, was relevantes Wissen ist und welche Fertigkeiten und Fähigkeiten Lernende weiterhin benötigen und welche vielleicht an Relevanz verloren haben.

6.2 Zum Bildungsbegriff

Oberstes Ziel von Bildung ist es, den Menschen zu einer mündigen und freien, das heißt einer zur Verantwortung fähigen Person heranzubilden. Bildungsbegriffe und -konzepte sind zwar immer auch vom jeweiligen Kulturkreis abhängig, insofern die Sicht auf den Menschen, seine Sozialität und damit die Verhältnisbestimmung von Individuum und Gesellschaft auch das Verständnis von Bildung prägt. Die folgenden Überlegungen gehen davon aus, dass Bildung sich an einem Verständnis des *humanum* orientiert, das auf eine umfassende Persönlichkeitsbildung zielt, die auf Urteilsfähigkeit und verantwortliche Teilhabe an der Gesellschaft angelegt ist.

Solche Bildung erfordert den Erwerb von Orientierungswissen als Bedingung reflexiver Urteilskraft und Entscheidungsstärke. Orientierungswissen entsteht einerseits aus der

218 Da die inhaltliche Befassung mit der Stellungnahme zum Zeitpunkt der Veröffentlichung von ChatGPT bereits abgeschlossen war, wird hier nicht detaillierter auf ChatGPT und die damit verbundenen Folgen für schulische Bildung eingegangen.

kritischen Rezeption empirischer Inhalte – von Alltagserfahrungen bis zu Ergebnissen wissenschaftlicher Forschung – und andererseits aus der Wahrnehmung normativer Inhalte – von sozialen Üblichkeiten bis zur reflektierten moralischen Urteilsbildung. Aus beiden Quellen bildet sich die Fähigkeit zur Übernahme von Verantwortung des Menschen gegenüber seiner sozialen Umgebung und sich selbst.

So umfasst Bildung neben der Vermittlung von Informationen ebenso Kontextwissen, das technische Kompetenzen vor dem Hintergrund verschiedener Erfahrungen zu beurteilen und einzusetzen vermag. Zu diesen Erfahrungen gehören politische, kulturelle oder existenzielle Erfahrungen, die in einem kulturellen Lernen intergenerationell weitergegeben und angeeignet werden können. Wenn über die Kriterien der Bildung angesichts voranschreitender Digitalisierung auch im schulischen Unterricht nachgedacht wird, ist das aus empirischen wie normativen Aspekten zusammengesetzte Orientierungswissens als Grundlage kulturellen Lernens zu berücksichtigen.[219] Es ist als Kriterium von Bildung auch für die Chancen und Risiken digitaler Bildung relevant.

Die Digitalisierung ist nicht Selbstzweck und die mit der digitalen Bildung oft einhergehende Vorstellung der Optimierung von Lernprozessen (oder des Lehr- und Lerngeschehens) ist kritisch auf das grundlegende Bildungsziel urteilsfähiger Selbstbestimmung des Menschen und seiner Handlungsverantwortung zu überprüfen.

Bildung, die auch im Rahmen von digitaler Bildung als Bildung zum Humanum zu begreifen ist, geht einher mit Identitätsstiftung. Aus den Überlegungen der vorangegangenen Kapitel ergibt sich, dass in diesem Feld der KI mögliche Lehrroboter eine Bildung, die sich als Kompetenz zur Selbstbestimmung und Urteilsfähigkeit auszeichnet, nicht ersetzen können. Sie können dem Kompetenzerwerb sogar abträglich sein, denn zum hier vertretenen Bildungsverständnis gehört

219 Nicht weiter eingegangen wird auf interkulturelles Lernen.

es, Ideen zu entwickeln, Geschichten zu entwerfen, Kunstwerke zu entdecken, aber auch Kritikfähigkeiten auszubilden, um Ideologien zu durchschauen. Das abwägende und kompetente Urteil der mündigen Person bildet hier die Voraussetzung für die Beurteilung der digitalen Prozesse.

Zu beachten ist daher ferner: Bildung erschöpft sich nicht in einer kognitiven Vermittlung von Information, sondern sie schließt affektive Dimensionen und soziale Kontakte ein. Schulisches Lernen ist daher nicht auf kognitiv-technische Operationen zu reduzieren, sondern beinhaltet auch emotionale und motivationale Aspekte. Für die Lehrkräfte ist es von Bedeutung, sich klarzumachen, dass das Lehr- und Lerngeschehen als dynamische Interaktion mit anderen Personen zu begreifen ist, als Interaktion mit Lehrenden wie auch mit anderen Lernenden. Lehrkräfte können im besten Fall als Identifikationsfiguren ein motivierendes und identitätsstiftendes Potenzial entfalten, das den Wissensdurst in Schülerinnen und Schülern wecken und Begeisterungsfähigkeit entfachen kann. Der Informationsgewinn durch Vermittlung von Sachwissen wird damit reflexiv auf die eigene Lebenswirklichkeit bezogen. Solches reflexive Lernen ist entscheidend für die Vorgänge des Verstehens und Aneignens.

Digitale Bildung ist daraufhin zu überprüfen, ob sie solch reflexives Lernen in der Interaktion zwischen Lehrkräften und Lernenden fördert oder behindert. Denn erst in diesem reflexiven Aneignen von Wissen werden die Potenziale jener Bildung freigesetzt, die den Menschen als freie und selbstverantwortliche Person ansprechbar werden lassen. Diese Kompetenzen reflexiven Verstehens und Aneignens sind daher Voraussetzung für einen urteilsfähigen Umgang mit den Möglichkeiten digitalen Lernens.

Bildung umfasst damit mehr als Informationszuwachs und logisches Schließen. Dazu zählt die immer mitlaufende Bezogenheit auf soziale und kommunikative Verständigung. Werden Teile der Interaktion zwischen Lernenden und Lehrenden oder auch der Lernenden untereinander durch digitale

Angebote ersetzt oder vermittelt, stellt sich die Frage, was dies für individuelle und soziale Lernprozesse bedeutet. Einer Automatisierung von Lernprozessen, die den reflexiven Bezug der Aneignung von Wissen und des Verständnisses erschwert oder gar unterbindet, wäre daher eine Absage zu erteilen. Damit würde das Lernen nicht mehr im vorgestellten umfassenden Sinne der Bildung der Person erfolgen. Ob Bildung sich ohne ein nicht zu vernachlässigendes personales Gegenüber vollziehen kann, ist fraglich. Denn für den Aneignungs- und Verstehensprozess ist der kommunikative Aspekt des Lernens, der ein personales Gegenüber voraussetzt, unverzichtbar.

Der Eingang der Digitalisierung in den Bereich der Bildung muss hinsichtlich der Veränderungsprozesse, die er für die Bildung als solche einleitet, daher genau reflektiert werden. Das gilt sowohl für Prozesse aufseiten der Lehrenden wie aufseiten der Lernenden. Dafür ist es zum Beispiel nötig, dass Lehrkräfte auch in der Lage sind, die Funktionsweise datenbasierter, KI-gestützter Lehr- und Lernsoftware hinsichtlich ihrer didaktischen Leistungsfähigkeit einzuordnen, um ihren Einsatz konstruktiv vornehmen, aber auch kritisch begleiten zu können. Es geht mit Blick auf die Zukunft darum, den Einsatz der neuen Techniken an den Entwicklungspotenzialen der Lernenden zu orientieren. Das Ziel muss sein, die Bildung zur Verantwortungsfähigkeit nicht zu vermindern, sondern zu erweitern. Aufgrund der Dynamik und Komplexität der Entwicklung solcher digitalen Lerntechniken ist es daher erforderlich, möglichst bald entsprechend der Zertifizierung klassischer Lehr- und Lernmaterialien auch deren digitale Pendants zu zertifizieren.[220]

220 Für die österreichischen Bemühungen vgl. Bundesministerium für Klimaschutz, Umwelt, Energie, Mobilität, Innovation und Technologie 2021, 30. Es werden allgemeine Zertifizierungskriterien entwickelt, die dann auch für die Zertifizierung des Einsatzes im schulischen Alltag Anwendung finden könnten (Heesen et al. 2020).

6.3 Einsatzmöglichkeiten datenbasierter, KI-gestützter Lehr- und Lernsysteme

Grundsätzlich eröffnet die Entwicklung zunehmend leistungsfähiger und kostengünstiger Geräte, die mobil verwendet, vernetzt und mit potenziell vielfältigen Sensoren ausgestattet oder verknüpft werden können, eine Fülle an (neuen) Einsatzmöglichkeiten im Lernraum Schule. Sie decken ein Spektrum ab, das von der Erfassung relevanter Daten und dem Erkennen von Mustern über deren Analyse zur Diagnostik von zum Beispiel Wissensständen und Lernfortschritten bis hin zu didaktischen Interventionen reicht.[221] Ausgangspunkt ist immer die Sammlung und Auswertung vieler Daten der Lernenden und mitunter auch der Lehrenden. Qualität und Umfang der verfügbaren Daten haben einen zentralen Einfluss auf die Leistungsfähigkeit von KI-Systemen.[222] Für Lehr- und Lernsysteme bedeutet das, dass die Genauigkeit von Prognosen über den Lernverlauf mit der Menge der verfügbaren Daten steigt.[223] Jede intensive – und mitunter auch invasive – Datenerhebung für den Einsatz von KI-gestützter Lehr- und Lernsoftware sollte je nach Zweck und Ziel des Einsatzes beurteilt und auf ihr sachlich angemessenes und ethisch vertretbares Maß überprüft werden. Hier stellen sich Fragen nach dem sinnvollen Grad und Ausmaß der Datenerhebung sowie deren wünschenswerten Verwendungsweisen. Dabei geht es beispielsweise darum, wie Datenerfassung die Lernenden in ihrem individuellen Lernprozess bestmöglich unterstützt, ohne dass diese Daten zur Überwachung oder Stigmatisierung von einzelnen Lernenden missbraucht werden können. Hier ist ein ethischer Umgang

221 Molenaar 2021, 62.
222 Bitkom 2017, 66 ff.
223 Martini et al. 2020, 23.

mit Daten auch im schulischen Bereich gefordert, der größtmögliche Datensouveränität auch für Schülerinnen und Schüler ermöglicht.[224]

Auf Grundlage der erhobenen Daten können individualisierte Rückmeldungen über Lehr- und Lernprozesse sowie entsprechende Reaktionen oder Empfehlungen des Softwaresystems erfolgen, die in der Folge die Grundlage für didaktische wie auch pädagogische Interventionen und auch Innovation bilden können. Dabei kommen zunehmend Elemente maschinellen Lernens zum Einsatz, bei denen die fortlaufende Analyse der Eingabedaten dafür genutzt wird, die diagnostischen und gegebenenfalls prädiktiven Ausgaben des Softwaresystems wie auch seine Empfehlungen dynamisch anzupassen. So können die Systeme durch Auswertung von beispielsweise Lerngeschwindigkeit, typischen Fehlern, Stärken und Schwächen das Lernprofil der Lernenden erkennen und die Lerninhalte entsprechend anpassen.[225] Die subjektiven Eindrücke der Lehrkräfte zu Lernfortschritt und Aufmerksamkeitsspanne individueller Lernender können dadurch datenbasiert untermauert, aber auch korrigiert werden.

Der Einsatz dieser Systeme kann dabei je nach Ziel, Gestaltung und technischen Grundlagen sehr unterschiedlich ausfallen. Er reicht beispielsweise von Programmen für

---

224 Um den Datenumgang verantwortlich zu gestalten, sind Grundkenntnisse über die Bedeutung und den Wert von Big Data und die damit verbundenen Risiken vorausgesetzt. Das heißt, der kritische Umgang mit Daten erfordert Datensouveränität. Dies hat der Deutsche Ethikrat bereits in seiner Stellungnahme zu Big Data von 2017 (271 f.) festgehalten: „Da bereits Kinder digitale Anwendungen nutzen und dabei Daten generieren, sollte eine entsprechende Nutzerkompetenz schon in der Schule vermittelt werden. Über die rein technischen Aspekte der gängigen Digitalisierungsstrategien schulischen Unterrichts hinaus sollte dies als Querschnittsaufgabe für alle Fächer des schulischen Curriculums ausgestaltet sein, um der gerade bei Kindern und Jugendlichen virulenten informationellen Selbstgefährdung entgegenzuwirken und schon früh ein Bewusstsein für die rechtlichen, sozialen und ethischen Implikationen zu schaffen. Die Vermittlung solcher Nutzerkompetenz sollte daher zukünftig Teil der Lehreraus- und -fortbildung werden."

225 https://www.zeit.de/digital/internet/2016-03/babbel-sprachkurs-start-up-berlin [11.01.2023].

personalisiertes Lernen auf individuellen Endgeräten, die den Lernenden Wissen vermitteln oder bestimmte Fähigkeiten trainieren (adaptive Lernprogramme), über die Erfassung und Rückmeldung bestimmter Fortschritte, Schwächen und Stärken der Lernenden an Lehrkräfte, um individuelle Förderung zu erleichtern (Lernstandkontrolle), bis hin zur Überwachung der Aufmerksamkeit oder gar der Emotionen der Lernenden und Lehrenden durch Sensoren und Kameras. Hinzu kommen Anwendungen aus dem Bereich der Robotik, Virtual Reality oder Augmented Reality. Zu manchen Entwicklungen in diesen Bereichen gibt es erhebliche Kontroversen (vgl. Abschnitt 6.4).

Häufig enthalten Lernprogramme spielerische Elemente (sogenannte Gamification), um die Motivation zur Nutzung und damit zum Lernen zu steigern.[226] Durch Nudging sollen die Lernenden angeregt werden, das *Richtige* zu tun, das heißt das gewünschte Lernverhalten auszuüben. Die Frage nach dem *richtigen* Lernen hat bereits Anlass zu Diskussionen dahingehend gegeben, ob Kinder und Jugendliche dadurch möglicherweise in ein „behavioristisches Dressursetting" hineingezogen werden.[227] Auch stellt sich die Frage, wer festlegt, was als *richtig* gelten darf. Im hier verwendeten Menschen- wie Bildungsbegriff ist bereits angelegt, an dieser Stelle kritisch gegenüber einem hohen Maß an Gamification und Nudging im Bildungsprozess zu sein. Auch hier sollte die Zielrichtung jedweden Einsatzes in der Erweiterung menschlicher Handlungsmöglichkeiten und Autorschaft liegen.

Auch wenn soziale Lernprozesse nicht durch KI ersetzt werden, werden soziale Beziehungen in Lernkontexten zunehmend durch KI vermittelt, beispielsweise durch die Messung von Leistung oder aber auch von Aufmerksamkeit oder Emotionen, deren Ergebnisse dann an Lehrkräfte und Lernende zurückgespiegelt wird. So können einerseits Informationen

226 Rachels/Rockinson-Szapkiw 2018.
227 Hartong 2019, 12.

über die Gruppe, zum Beispiel eine Schulklasse, ausgewertet werden, um daraus Hinweise auf die Lerngeschwindigkeit und inhaltlichen Schwächen abzuleiten, mittels derer die Lehrkräfte ihre kurz-, mittel- und langfristigen Strategien besser an die jeweilige Gruppe anpassen können. Auf der individuellen Ebene können KI-gestützte Systeme dazu beitragen, individuelle Schwächen oder Rückstände zu identifizieren oder gar die Gefahr des Schulabbruchs frühzeitig anhand von Mustern zu erkennen und gezielte Unterstützungsangebote zu unterbreiten. Auf die Risiken solcher Ansätze wird im Folgenden noch eingegangen.

6.4 Mensch-Maschine-Relationen in der schulischen Bildung: Erweitern, Vermindern, Ersetzen

Ähnlich wie im Medizinbereich muss auch im Bildungsprozess von unterschiedlichen Beziehungen ausgegangen werden, die durch den KI-Einsatz betroffen sind: Interaktionen zwischen Lehrenden und Lernenden wie auch der Lernenden untereinander im Klassenraum, im weiteren Sozialraum Schule, aber auch darüber hinaus. Hinzu kommen mitunter spezifische Untergruppen wie Lernende mit besonderen Bedürfnissen und gegebenenfalls weitere Beteiligte wie zusätzliches (sozial-)pädagogisches Personal, Eltern, Personen in der Schulverwaltung sowie jene, die an der Entwicklung und Planung von Curricula und dem Einsatz digitaler Techniken beteiligt sind. Parallelen zum Handlungsfeld Medizin lassen sich auch insofern feststellen, als es in der Schulbildung ebenfalls zu engen, mittleren und weitgehenden Ersetzungen bestimmter Handlungssegmente und Interaktionen kommen kann, die sich jeweils situationsabhängig erweiternd oder vermindernd auf die Handlungsmöglichkeiten der Beteiligten auswirken können.

Im Folgenden werden exemplarisch einige KI-basierte Lehr- und Lernsysteme für den schulischen Kontext dargestellt. Die

genannten Möglichkeiten durch KI im System Schule sollten vor dem Hintergrund des zuvor dargelegten Menschenbildes und Bildungsbegriffs bewertet werden. Hier gilt insbesondere, dass die für Bildungsprozesse und Persönlichkeitsentwicklung grundlegenden Aspekte der direkten Ansprache sowie der personalen Beziehungen nicht vernachlässigt werden dürfen. Mit Blick auf sehr weitreichende Ersetzungen ist dabei auch auf die sich stellende Gefahr des Animismus, das heißt der Zuschreibung mentaler Eigenschaften an Maschinen zu verweisen, die bei sich noch entwickelnden Kindern und Jugendlichen umso schwerer wiegt. Engere und gegebenenfalls mittlere Ersetzungen können aber durchaus mit den normativen Weichenstellungen vereinbar sein und den Bildungsprozess unterstützen.

Eine *enge Ersetzung* liegt etwa vor, wenn ein Softwaresystem für einen genau bestimmten Lernabschnitt wie das Vokabellernen im Fremdsprachenunterricht oder das Formellernen in der Mathematik eingesetzt wird. Ein Beispiel für solch ein eng umrissenes algorithmisches System, dass beim Erlernen einer ganz spezifischen Kompetenz zum schulischen Einsatz kommt, ist das intelligente Tutorsystem Subkraki[228], mit dessen Hilfe die Lernenden in der Grundschule das schriftliche Subtrahieren üben können. Das Programm stellt Übungsaufgaben, erkennt, welche Art von Fehlern die Lernenden bei der Lösung der gestellten Aufgabe machen, und gibt dann konstruktives Feedback, indem es beispielsweise spezifisch auf das einem beobachteten Fehler zugrundeliegende Verständnisproblem reagiert. Fehlerhafte Lösungen werden nicht einfach vom System korrigiert. Stattdessen werden die Lösungsschritte anhand einer ähnlichen Rechenaufgabe erläutert und den Lernenden damit die Möglichkeit gegeben, Fehler eigenständig zu korrigieren, daraus zu lernen und Selbstwirksamkeit zu entwickeln.

228 Zeller/Schmid 2017. Für eine Demo-Version von Subkraki siehe https://cogsys.uni-bamberg.de/ITS [02.03.2023].

Aufwendigere und datenintensivere intelligente Tutorsysteme (siehe Infokasten 4) können auch komplexere Lerninhalte in verschiedenen Fächern im Zusammenwirken mit Lernenden vermitteln und so breitere Teilaspekte des Unterrichtsgeschehens ersetzen (mittlere Ersetzung) oder im Einzelfall die Funktion einer Lehrkraft vollständig übernehmen (vollständige Ersetzung). Eine weitgehende Ersetzung kann noch zugespitzt werden, wenn ein als Software realisiertes KI-System in bestimmten Situationen nicht nur als eigenständige Lehrkraft auftritt, sondern über einen virtuellen Avatar oder in einem Roboter mit besonders personalen Eigenschaften wie Mimik oder natürlicher Sprache in Erscheinung tritt.

Infokasten 4: Intelligente Tutorsysteme

Schon in den 1970er-Jahren wurden intelligente Tutorsysteme (ITS) zur Unterstützung individueller Lernprozesse entwickelt. Aktuelle Fortschritte bei datengetriebenen, KI-gestützten Ansätzen haben die Möglichkeiten solcher Softwaresysteme deutlich erweitert. Ziel von ITS ist es, Inhalte auf eine Art und Weise zu vermitteln, die sich dynamisch an die individuellen Besonderheiten und Bedürfnisse der Lernenden anpasst.

ITS bestehen typischerweise aus vier Komponenten[229]: Sein Kern ist (1) ein Wissensmodell, das sämtliche Konzepte, Regeln und Problemlösungsstrategien für das zu vermittelnde Wissen umfasst. Aufgabenstellungen können vom Wissensmodell eigenständig gelöst werden. Es entspricht also einem Expertensystem für den zu vermittelnden Wissensbereich – sei es analytische Geometrie, Hebelgesetze oder die Entstehung von Regen. Das Computerprogramm gleicht dieses Wissensmodell fortlaufend mit den verfügbaren Daten der Lernenden ab, die es in (2) einem Studierendenmodell zusammenfasst. Dabei analysiert das System, wie Lernende sich zum Wissensmodell verhalten – welche Fortschritte oder Fehler sie beispielsweise machen – und aktualisiert das Studierendenmodell immer wieder entsprechend. Im einfachsten Fall ist dies ein Fortschrittsprofil. Es können aber auch komplexe intelligente Methoden zum Einsatz kommen, die die Identifikation von Fehlkonzepten ermöglichen. In (3) der Didaktikkomponente wählt das System auf Grundlage dieses Abgleichs bestimmte Rückmeldungen an die Lernenden (sowie gegebenenfalls auch an Lehrkräfte) aus, etwa Einschätzungen zum Fortschritt, Hinweise auf Fehler und Hilfestellungen zu ihrer Überwindung oder eine Auswahl weiterführender Schritte oder Aufgaben. An (4) der Interaktionsschnittstelle schließlich findet der Austausch zwischen dem Tutorsystem und den Lernenden statt, beispielsweise in Form von Texteingaben und -ausgaben.

229 Nkambou/Bourdeau/Mizoguchi 2010, 3 ff.

Während in frühen ITS sämtliche möglichen Bezüge zwischen dem Wissens- und dem Studierendenmodell ebenso wie alle Ausgabemöglichkeiten vorab bestimmt und programmiert werden mussten, zum Beispiel in Form von Entscheidungsbäumen, erlauben moderne datenintensive und KI-gestützte Ansätze flexiblere und differenziertere Funktionen. Je nach Menge und Vielfalt der Daten, die Lernende in das System einspeisen, können der individuelle Wissensstand sowie Fortschritte und gegebenenfalls auch Präferenzen beim Lernen genauer analysiert werden. Dadurch kann das Programm präzisere Vorhersagen für den weiteren Lernprozess treffen und seine Ausgaben entsprechend individualisieren.

Mit ihren vier Komponenten bilden ITS einige Kernfunktionen nach, die sonst von menschlichen Lehrkräften übernommen werden: (1) die Bereithaltung von Fachwissen, (2) die Einschätzung des Lernstands und Förderbedarfs von Lernenden, (3) eine daran angepasste Selektion didaktischer Strategien und (4) die Interaktion mit den Lernenden. ITS können daher mit dem Anspruch eingesetzt werden, die Funktion einer Lehrkraft für den inhaltlich abgedeckten Bereich ganz oder teilweise zu ersetzen. Genauso ist jedoch ein Einsatz in hybriden Formaten und/oder mit Schwerpunkten bei der Unterstützung der Lehrkraft möglich.

Als Beispiel für ein System mit weitreichendem Ersetzungspotenzial kann AutoTutor genannt werden. Es wurde von Arthur Graesser am Institute for Intelligent Systems der University of Memphis entwickelt. Das Programm simuliert eine menschliche Lehrkraft, indem es mit den Lernenden ein natürliches Gespräch führt.[230] Dabei erfolgt eine Anpassung an die Lernenden, indem ihre Emotionen und Reaktionen (z. B. Dialogmuster, Mimik, Körperhaltung) erfasst werden. Bisher mit dieser Methode unterstützte Inhalte aus dem schulischen und außerschulischen Bereich stammen dabei unter anderem aus der Informatik, Physik, Mathematik, Elektronik, Philosophie und Biologie. AutoTutor bietet darüber hinaus die Möglichkeit, jeweils die Über- und Unterforderung der Lernenden zu registrieren und darauf zu reagieren. Auf diese Weise eingesetzt, können ITS dazu beitragen, positive Lernumgebungen zu schaffen. ITS bieten die Möglichkeit, Lerninhalte von Mathematik bis Fremdsprachenerwerb in spezifischen Aufgabenkontexten zu üben und dabei unmittelbares und individualisiertes Feedback zu erhalten. Somit sind sie geeignet, den Unterricht mit einer Lehrkraft im Klassenverband gezielt um Phasen des individualisierten Lernens zu ergänzen. Kompetenzen werden konstruktiv (Learning by Doing) erworben und ergänzen das *learning by being told* des üblichen Unterrichts.

Intelligente Tutorsysteme und verwandte daten- und KI-gestützte adaptive Lerntechnologien werden zunehmend weniger mit dem Ziel entwickelt, eine menschliche Lehrkraft möglichst weitgehend zu ersetzen, sondern immer stärker für den Einsatz in hybriden Lehr- und Lernsystemen konzipiert.[231]

230 Graesser et al. 2012.
231 Molenaar 2021.

Solche Systeme werden in zahlreichen Ländern bereits mit gut dokumentierten Beiträgen zur Verbesserung von Lernerfolgen bei unterschiedlichen Altersgruppen eingesetzt. Das 2012 aus einer Elterninitiative in den Niederlanden hervorgegangene Angebot Snappet beispielsweise wird inzwischen in mehreren Ländern, darunter auch Deutschland, zur Unterstützung des Unterrichts in der Grundschule verwendet. Die Software wird von Lernenden im Präsenz-, Distanz- oder Wechselunterricht zur Bearbeitung von Aufgaben eingesetzt, die auf die Lehrpläne und den individuellen Lernfortschritt in allen Fächern zugeschnitten sind. Sie spielt die Auswertung der dabei gesammelten Daten an die Lehrkräfte zurück, um diese bei der weiteren Unterrichtsplanung und differenzierten Förderung der Kinder zu unterstützen. In einer Studie an 79 Grundschulen konnten mit dem Einsatz von Snappet nach fünf Monaten unter anderem in einer standardisierten Prüfung verbesserte Ergebnisse und eine höhere intrinsische Motivation der Lernenden nachgewiesen werden.[232]

Auch eine weiter reichende Nutzung von Technik zur erfolgreichen Teilhabe am Schulunterricht kann im Einzelfall sinnvoll sein. Dies kann mit einem Szenario veranschaulicht werden, in dem ein Telepräsenzroboter Lernenden, die nur von zu Hause am Unterricht teilnehmen können, mehr Interaktion mit der Lerngruppe ermöglicht (siehe Infokasten 5).

Infokasten 5: Teilnahme am Unterricht mittels eines Telepräsenzroboters

Mittels eines Telepräsenzroboters (z. B. AV1 Avatar des norwegischen Startups No Isolation) können Lernende im Falle längerer Abwesenheiten, beispielsweise aufgrund von Krankheit, ihre Klasse virtuell besuchen. Der Roboter übernimmt den Platz des Lernenden in der Klasse, ausgestattet mit Internetverbindung, Kamera, Lautsprecher, Mikrofon. Die Lernenden können damit im Klassenzimmer in Echtzeit hören, sehen und sprechen (via App am Smartphone oder Tablet von zu Hause aus, ohne Datenspeicherung). Die Funktionen des Roboters umfassen das Herumblicken im Klassenzimmer,

232 Faber/Luyten/Visscher 2017.

das Handheben und Sprechen oder auch Mimik, um zum Beispiel Unverständnis auszudrücken.[233]

Soziale Isolation zum Beispiel infolge einer chronischen Erkrankung kann durch diesen Telepräsenzroboter abgemildert, eine Wiedereingliederung unterstützt beziehungsweise der Platz im Klassenzimmer symbolisch vom Telepräsenzroboter besetzt gehalten werden. Der Telepräsenzroboter unterstützt so nicht nur kognitive, sondern in beschränktem Ausmaß auch emotionale und soziale Funktionen, indem er beispielsweise das Tuscheln mit anderen Kindern aus der Klasse ermöglicht. Es geht also nicht nur darum, im Lernpensum nicht zurückzubleiben, sondern auch darum, weiterhin als Teil der Klassengemeinschaft wahrgenommen zu werden. Während der AV1-Roboter menschenähnliche Züge aufweist, wird hierauf in anderen Projekten explizit verzichtet, um anthropomorphen Fehlverständnissen vorzubeugen.

Neben den zuvor beschriebenen Möglichkeiten für den Einsatz datenbasierter, KI-gestützter Technologien in der personalisierten Gestaltung und Auswertung konkreter Lerninhalte und -prozesse gibt es mittlerweile auch Bestrebungen, KI zur Analyse des Verhaltens im Klassenraum oder ganzer Einrichtungen einzusetzen. Wie umfassend dies geschehen kann, zeigen Beispiele aus China (siehe Infokasten 6).

Infokasten 6: Besonders weitreichender Einsatz digitaler Technologien im Schulunterricht

Teilweise kann ein besonders weitreichender Einsatz von datenintensiven und KI-gestützten Tools im Schulunterricht beobachtet werden, der über die Verwendung einer geschlossenen Lehr- und Lernsoftware hinausgeht.[234] So werden beispielsweise an manchen chinesischen Schulen mithilfe zahlreicher und vielfältiger Sensoren umfangreiche Daten auf dem Campus gesammelt. Dies geschieht einerseits, um die Ressourcen auf dem Gelände zu verwalten. Dazu gehören unter anderem die Sicherheitseinrichtungen, die Beleuchtung, die Regelung der Wasser- und Luftqualität, aber auch das Sammeln von Bewegungsdaten. Im Klassenzimmer erfolgt eine ähnlich umfassende Datensammlung, beispielsweise über Kameras und Mikrofone und teilweise ergänzt um individuelle, von den Lernenden getragene Geräte (Wearables), um ein möglichst umfassendes Bild eines jeden Individuums zu erstellen. Es werden Körperdaten (z. B. Herzfrequenz, Körpertemperatur) ermittelt, aber auch Daten zum Sozialverhalten, zur emotionalen und psychischen Gesundheit anhand von Analysen der Mimik, Körperhaltung und

233 Belpaeme/Tanaka 2021, 148.
234 Vincent-Lancrin 2021, 24 f.

Sprache sowie natürlich zu schulischen Leistungen. Genannte Ziele dieser Maßnahmen umfassen die Sicherstellung der Aufmerksamkeit der Schülerinnen und Schüler, die individuelle Förderung der Lernenden sowie die Unterstützung der Lehrkräfte, beispielsweise bei Aufgaben wie der Unterrichtsvorbereitung, dem Vergeben von Hausaufgaben und der Auswertung von Lernprozessen.

Ein Schwerpunkt solcher Ansätze wird mit der Bezeichnung „Classroom Analytics" beschrieben. Damit sind Konzepte gemeint, welche die Dynamik ganzer Lerngruppen umfassend zu dokumentieren und auszuwerten versuchen, unter Berücksichtigung des Verhaltens und der Interaktion der Lernenden und Lehrkräfte im Klassenraum[235], einschließlich der räumlichen Gegebenheiten und Prozesse vor Ort[236]. Classroom-Analytics-Ansätze sind aufgrund der für sie notwendigen Erfassung vielfältiger Daten unter anderem über das Verhalten von Schülerinnen und Schülern sowie Lehrkräften umstritten (siehe unten). Befürworter solcher Ansätze verweisen auf die Chancen, die sich durch Classroom Analytics hinsichtlich einer Verbesserung von Pädagogik und Didaktik ergeben können. Beispielsweise könne Classroom Analytics die Kapazität von Lehrkräften steigern, alle Lernenden gleichzeitig im Blick zu haben und sie so im gesamten Management ihrer Lehrkonzeption zu unterstützen, etwa bei der Bildung von Lerngruppen oder bei der Frage, wann in die Lernprozesse der Schülerinnen und Schüler eingegriffen werden soll.

Mittels Classroom Analytics können der Lehrkraft Informationen an die Hand gegeben werden, die auch zur kritischen Selbstreflexion des Lehrverhaltens genutzt werden können. Für computergestützte Analysen ist beispielsweise das Verhalten des lehrenden Fachpersonals schwierig zu erfassen, wenn nur die verbale Kommunikation zugrunde gelegt wird. Indem jedoch mittels Classroom Analytics auch das Verhalten im Klassenraum beobachtet wird, zu dem auch die Häufigkeit

235 Dillenbourg 2021.
236 Martinez-Maldonado et al. 2022.

der persönlichen Kontakte mit den Schülerinnen und Schülern gehört, kommen zusätzliche Parameter ins Spiel, die pädagogische Ansätze und Lehrstrategien messbarer werden lassen. Damit kann den Lehrkräften in konstruktiver Weise ein Feedback gegeben werden. Ein solches Instrument ist etwa das TeachActive-Dashboard. Es kann die pädagogischen Bemühungen durch visuelle Unterstützung verbessern. Den Lehrkräften wird mit den Analysen gespiegelt, ob und wie ihre pädagogischen Praktiken zum aktiven Lernerfolg der Lernenden beitragen.[237] Classroom Analytics eröffnen Lehrkräften zudem Chancen auf eine objektivere Wahrnehmung der Lernenden und können es ermöglichen, implizite Vorurteile der Lehrkräfte gegenüber Lernenden, etwa zu Herkunft und Gender, aufzudecken.[238] Wenn solches Feedback den Lehrkräften zurückgespiegelt wird, kann dies als kritisches Korrektiv für das eigene Handeln dienen.

In diesem Zusammenhang sind unterschiedliche ethische Fragen relevant. Da im Rahmen von Classroom Analytics mitunter besonders umfangreiche Daten erfasst und ausgewertet werden, sind zum einen negative Auswirkungen solcher Messungen auf die Privatsphäre und Autonomie aller Beteiligten zu diskutieren. Neben den direkten Auswirkungen der Datenerfassung und Prognostik auf die Privatsphäre und Autonomie der Beobachteten ist hier insbesondere auch auf sogenannte Chilling-Effekte hinzuweisen. Im Kontext von Überwachung beschreiben Chilling-Effekte das Phänomen, dass bereits die Sorge, bestimmte Daten über Personen könnten erfasst und ihr Verhalten analysiert werden, negative Rückwirkungen auf ihr Befinden und Verhalten haben kann. Dies könnte eine Verminderung von Handlungsfreiheit, Motivation und anderen für gute Bildung wichtigen Bedingungen mit sich bringen. Darüber hinaus können Daten in unangemessener Weise zweck- und kontextentfremdet genutzt werden, wenn zum

237 Alzoubi et al. 2021.
238 Reinholz/Stone-Johnstone/Shah 2020.

Beispiel Informationen über mangelhafte Schulleistungen oder Lehrkraftevaluierungen zu Strafen oder anderweitigen Sanktionen führen. Ein weiteres Problem stellen systematische Verzerrungen (Bias-Problematik) sowohl in der Datenanalyse als auch in der Interpretation von Daten und Prognosen dar, welche zu ungerechten Urteilen über die Leistung von Lernenden oder Lehrenden führen können.

Um diesen Herausforderungen zu begegnen, bedarf es eines verantwortlichen Umgangs mit Daten und Analysen, der die Bedürfnisse von Lernenden und Lehrenden stets im Blick hat. Diese Personengruppen sollten bereits während der Entwicklung digitaler Bildungsinstrumente mit einbezogen werden. Die ethische Angemessenheit von Lösungen hängt dabei auch immer von technischen und institutionellen Details ab: welche Daten über wen in welcher Granularität erfasst werden (z. B. individuell, gruppen- oder aufgabenbezogen) und wer in welcher Granularität und für welche Zwecke Zugang zu den Daten, Analysen oder Prognosen erhält.

Ein besonders kontrovers diskutierter Aspekt von Classroom Analytics betrifft in diesem Zusammenhang die mögliche Erfassung der Aufmerksamkeit (Attention Monitoring) oder der emotionalen Verfasstheit (Affect Recognition) der im Klassenraum interagierenden Personen, insbesondere basierend auf der Analyse von Audio- oder Videodaten. Die Bedeutung von Emotionen für Lernprozesse wurde in den letzten Jahrzehnten immer stärker berücksichtigt und entsprechend ist es nicht verwunderlich, dass nicht nur die Messung von Aufmerksamkeit, sondern auch von Emotionen in schulischen Kontexten voranschreitet. Die Verfügbarkeit von Sensoren wie Kameras und Mikrofonen in Laptops und Tablets, aber auch deren Einbau in Schulräumen scheinen die Möglichkeit der Erfassung von Aufmerksamkeit und Emotionen nahezulegen, um neben der Erfassung von Leistung zusätzliche Einsichten in das Lerngeschehen zu gewinnen. Auch wenn dies durchaus mit dem Ziel einer Verbesserung von Lernergebnissen verbunden sein kann, so ist der Einsatz solcher Technologien in

schulischen Kontexten aufgrund erkenntnistheoretischer und ethischer Herausforderungen kritisch zu betrachten.

Die *erkenntnistheoretische* Kritik insbesondere an Technologien zur Affekterkennung bezieht sich hier vor allem auf das vorherrschende reduktionistische Modell menschlicher Emotionen, nach dem es sechs Grundemotionen gibt, die als universell stabil gelten und die zuverlässig aus äußerem Verhalten wie insbesondere der Mimik erschlossen werden können. Der Kritik an diesem Modell wird mittlerweile in der Forschung durchaus Rechnung getragen, etwa mit der Erfassung situativer Kontexte oder der Zusammenführung verschiedener Datentypen, beispielsweise visueller und verbaler Daten mit dem Ziel, die Validität der Affektmessung zu erhöhen. Aber selbst wenn Emotionen in Zukunft zuverlässig durch KI-Systeme erfasst werden könnten, so sind dennoch *ethische* Bedenken hinsichtlich ihres Einsatzes in schulischen Kontexten zu berücksichtigen, da das hierfür notwendige Monitoring von Lernenden und Lehrenden Auswirkungen auf deren Privatsphäre und Autonomie haben kann und Fragen der Gerechtigkeit berührt.

Die Bewertung des Einsatzes von Technologien zur Affekt- und Aufmerksamkeitserkennung in schulischen Kontexten hängt daher von der Beantwortung folgender Frage ab: Können und sollten Aufmerksamkeit und Emotionen genau, zuverlässig und ohne systematische Verzerrung gemessen werden? Ob sie messbar *sind*, hängt von der wissenschaftlichen und technologischen Basis der eingesetzten Systeme ab.

Ob solche Instrumente eingesetzt werden *sollen*, hängt jedoch von einer zusätzlichen Bewertung des potenziellen Nutzens und Schadens für alle Beteiligten ab. Erstens müsste nachgewiesen werden, dass das Monitoring von Aufmerksamkeit und Emotionen insbesondere den Schülerinnen und Schülern nützt, zum Beispiel indem es ihre Lernprozesse oder Lernergebnisse verbessert. Zweitens müsste bewertet werden, ob diese potenziellen Vorteile die potenziellen Nachteile überwiegen, wie etwa die negativen Auswirkungen der notwendigen

Datenerfassung auf die Privatsphäre, Autonomie und Freiheit der Lernenden. Dazu gehören auch Fragen zu Chilling-Effekten und den langfristigen Folgen des Messens und Spiegelns insbesondere von Emotionen für Lernende und Lehrende. Drittens ist es nicht nur ein erkenntnistheoretisches, sondern auch ein ethisches Problem, wenn Technologien zur Aufmerksamkeits- und Emotionserfassung ungenau oder verzerrt arbeiten. Denn sobald ihr Einsatz zu weiteren Handlungen beziehungsweise zur Bewertung der Leistung führt, werfen solche Ungenauigkeiten und Verzerrungen auch Fragen der Gerechtigkeit auf.

Der Kontext von Schulen bringt zwei zusätzliche Herausforderungen mit sich. Zum einen sind Schulkontexte durch Hierarchien und asymmetrische Machtverhältnisse gekennzeichnet. Zum anderen sind Kinder beteiligt, die besonders vulnerabel sind, weshalb höhere Anforderungen an den möglichen Nutzen solcher Technologien für die beteiligten Kinder zu stellen sind.

Auch wenn ein Einsatz aktuell verfügbarer Technologien im schulischen Kontext in Anbetracht der oben genannten Kriterien nicht empfehlenswert erscheint, ist nicht prinzipiell auszuschließen, dass zukünftige Entwicklungen zu einer didaktisch sinnvollen Verbesserung des Lernprozesses beitragen.

In Anbetracht der erkenntnistheoretischen und ethischen Herausforderungen und unter Abwägung potenzieller Nutzen und Schäden stehen die Mitglieder des Deutschen Ethikrates dem Einsatz von Audio- und Videomonitoring im Klassenzimmer insgesamt skeptisch gegenüber. Insbesondere erscheint die Analyse von Aufmerksamkeit und Emotionen per Audio- und Videoüberwachung des Klassenraums mittels aktuell verfügbarer Technologien nicht vertretbar. Ein Teil des Ethikrates schließt den Einsatz von Technologien zur Aufmerksamkeits- und Affekterkennung zukünftig jedoch nicht vollständig aus, sofern sichergestellt ist, dass die erfassten Daten eine wissenschaftlich nachweisbare Verbesserung des Lernprozesses bieten und das hierfür notwendige Monitoring

von Schülerinnen und Schülern sowie Lehrkräften keine inakzeptablen Auswirkungen auf deren Privatsphäre und Autonomie hat. Ein anderer Teil des Ethikrates hingegen befürwortet ein Verbot von Technologien zu Aufmerksamkeitsmonitoring und Affekterkennung in Schulen.

6.5 Grundsätzliche Diskussion: KI im schulischen Bildungsprozess

Die öffentliche Debatte zu KI in der Bildung hat sich in den letzten Jahren intensiviert und wird hauptsächlich in den Fach-Communitys und von Stakeholdern betrieben. So hat sich die Gewerkschaft Erziehung und Wissenschaft im Projekt „Bildung in der digitalen Welt" mit dem Thema Learning Analytics auseinandergesetzt.[239] Auch im Abschlussbericht der Enquete-Kommission Künstliche Intelligenz des Deutschen Bundestages wird das Potenzial von Learning-Analytics-Systemen hervorgehoben.[240]

Aufseiten der Chancen ist das personalisierte Lernen anzuführen, das im Sinne des hier skizzierten Feldes bedeutet, dass durch personalisierte Abstimmung auf Einzelne Lernchancen erhöht werden können. So können auf unterschiedliche Weise Stärken wie Schwächen des je einzelnen Lernenden wie auch einer Lerngruppe erkannt und auf existierende Schwierigkeiten beim Aufgabenlösen eingegangen werden. Dies kann einerseits Lernende unterstützen, sowohl in Bezug auf individuelle Stärken als auch besondere Bedürfnisse, da sie fokussiert in ihren Bedarfen von einem nicht müde oder ungeduldig werdenden System begleitet werden. Mithilfe dieser Prozesse werden auch Lehrkräfte unterstützt, die ihre Ressourcen dann sinnvoll im Gesamtlehrprozess einsetzen können. Als Beispiel sei das schon erwähnte Subkraki angeführt, das als eine Form

239 Hartong 2019.
240 Deutscher Bundestag 2020, 306.

der engen Ersetzung der Lehrenden das Lehr- und Lerngeschehen punktuell unterstützen kann und daher eine Entlastung der analogen Lehre für alle Beteiligten bietet. Eine weitere Chance besteht darin, dass KI-gestützte Lehr- und Lernsysteme für bestimmte Bevölkerungsgruppen die Zugangschancen zu Bildung verbessern können, selbst wenn sie nicht als Teil des Unterrichtsprogramms von der Schule angeboten werden. Auch wenn derartige Programme, etwa zum Spracherwerb oder zur Nachhilfe, meist nicht kostenfrei verfügbar sind, liegen die Kosten in der Regel weit unterhalb der Kosten für private Nachhilfe oder Sprachkurse.

Soweit mit dem Einsatz KI-gestützter Lehr- und Lernsysteme einzelne Lernende in den Fokus gerückt werden, ist ihr Einsatz im Besonderen auch unter dem Gesichtspunkt der Inklusion bedenkenswert. So können KI-basierte Lehr- und Lernsysteme für Lernende mit besonderen Bedürfnissen, deren individuelle Lerngeschichten adaptiver begleitet werden müssen, Chancen bieten.[241] KI kann bei der Entwicklung sozialer Fähigkeiten von autistischen Kindern helfen, Lernende mit Dysgrafie diagnostizieren und unterstützen sowie die Teilhabe für blinde oder sehbehinderte Kinder durch Bereitstellung grafischen Materials fördern.

Es bleibt deshalb grundsätzlich festzuhalten: Die neuen Systeme können hilfreiche, auf das Individuum zugeschnittene Einzeltools darstellen, sind aber nicht die generelle Lösung für Fragen von Inklusion, und sie müssen unter Berücksichtigung möglichen Missbrauchs und möglicher Risiken angewendet werden. Um diese zu erkennen und im Blick zu behalten, ist es wichtig, das digitale Lehr- und Lerngeschehen permanent zu evaluieren.

Eine weitere Chance, die der Einsatz technischer Systeme für Inklusion mit sich bringt, liegt in der Flexibilität und einem erweiterbaren Aktionsradius. Vor allem das Beispiel mit dem Telepräsenzroboter zeigt, wie Technik Inklusion erhöhen

241 Good 2021.

kann, indem sie es beispielsweise chronisch kranken Lernenden ermöglicht, sich von zu Hause aus am Unterrichtsgeschehen zu beteiligen.

Die Verwendung von Lehr- und Lerntechnologien ist also mit großen Chancen verbunden. Dazu gehört auch der berechtigte Wunsch nach einer objektiveren und faireren Bewertung der Lernergebnisse. Da Software nicht die möglichen Vorurteile der Lehrkräfte „hegt", sei es bezüglich der Intelligenz der Lernenden, seien es emotionale Vorbehalte, die eine Gleichbehandlung trüben können, steigt die Chance, dass Lernende mithilfe von KI-Anwendungen neutraler unterrichtet werden können. Allerdings darf auch hier eine solch vorgebliche Neutralität oder Objektivität nicht als gegeben angenommen werden, sondern bedarf der Überprüfung. Ein besonderes Risiko bei datenbasierten KI-Systemen besteht nämlich in systematischen Verzerrungen (Bias), woraus sich hohe Anforderungen an die Validität und Qualität der verwendeten Trainingsdaten und die Angemessenheit der verwendeten Methoden ergeben. Auch weitere allgemeine Risiken wie beispielsweise Fragen der Sicherheit, Privatheit und Transparenz zeigen sich natürlich ebenso beim Einsatz von KI in der Bildung.

Darüber hinaus ergeben sich im Kontext der (schulischen) Bildung weitere sektorspezifische Herausforderungen. Hier sind in besonderer Weise die sozialen Aspekte hervorzuheben, die beim Ersatz des Lehrpersonals eintreten können, insbesondere Gefahren der Isolation und Vereinsamung von Lernenden.

Außerdem ist zu berücksichtigen, dass bereits der Einsatz digitaler Medien das qualitative Ergebnis des Lernprozesses verändern kann. So werden Texte, die über elektronische Medien rezipiert werden, anders durchdrungen als Texte, die mittels analoger Medien (auf Papier) gelesen werden.[242] Auch beim Einsatz von KI muss in Betracht gezogen werden, dass sich Lernverhalten qualitativ verändert. So könnten sich etwa

242 E-READ 2019.

grundsätzliche Auswirkungen auf die Motivation und Fähigkeit von Schülerinnen und Schülern, komplexere Aufgaben zu lösen, ergeben. Wenn KI vor allem auf die Lösung kleiner, überschaubarer Aufgaben und schnelle Erfolgserlebnisse ausgerichtet ist, könnte langfristig die Bereitschaft vermindert werden, komplexere Probleme anzugehen, deren Lösung in der Regel nicht innerhalb kurzer Zeit gefunden werden kann.

Die Erforschung der Auswirkungen des Einsatzes digitaler Medien wie auch KI-basierter Methoden muss daher vielfältige, auch indirekte Dimensionen erfassen und darf sich nicht nur an wenigen offensichtlichen Markern schulischen Erfolgs (wie etwa den Schulnoten oder dem Abfragen von Fakten) orientieren. Sind diese Aspekte jedoch erst einmal hinreichend verstanden, so bieten sich gerade Methoden der KI an, um spezifische Defizite eines stärker digitalisierten Lernens zu kompensieren und etwa die individuelle Kompetenz zum Lösen komplexerer Aufgaben zu fördern.

Auch die Messung von Aufmerksamkeit und Emotionen kann ein Mittel sein, Lernprozesse und Lernerfolge besser zu verstehen. Allerdings sind Technologien der Aufmerksamkeits- und Emotionserkennung beispielsweise durch Kameras oder andere Instrumente wissenschaftlich umstritten[243] und greifen in invasiver Weise in die Autonomie und Privatheit der Lernenden ein. Eine solche invasive Form der Datenerfassung ist mit dem hier leitenden Bildungsbegriff, der sich am freien und urteilsfähigen Menschsein orientiert, nicht kompatibel und muss daher abgelehnt werden.[244]

---

243 Tcherkassof/Dupré 2022; Mohammad 2022.

244 Hansen et al. 2020; Hartong 2019. Der Einsatz umfangreicher Datenerhebungen und -auswertungen bringt vielfache Herausforderungen mit sich. Entsprechend spielen Datenschutzfragen und Sorgen vor einem möglichen Missbrauch von KI-gestützten Lehr- und Lernsystemen, vor allem bei flächendeckender Einführung, eine große Rolle im Fachdiskurs, ebenso wie eine mögliche Regulierung zur Verhinderung solcher Probleme. Im Vorschlag für einen Verhaltenskodex für Trusted Learning Analytics wird auf Kompetenzaufbau als Voraussetzung für einen verantwortungsvollen Umgang mit KI-gestützten Technologien gesetzt statt auf Regulierung.

Neben Fragen des Zugriffs auf Daten (z. B. durch Lehrkräfte, Eltern, Behörden) und dessen Auswirkungen auf Fragen der Privatsphäre müssen auch die Auswirkungen einer so engmaschigen Überwachung für die Lernatmosphäre und die Selbstwahrnehmung der Lernenden berücksichtigt werden. Hier spielen auch Fragen der Arbeitsplatzüberwachung eine Rolle, wenn auch Lehrpersonen in den Fokus der Technologie geraten. Hinzu kommen Zweifel an der Validität bestimmter Daten und Analysen. Diskutiert wird etwa, ob der Einsatz von EEG-Headsets und anderen Technologien zur Messung von Aufmerksamkeit oder emotionalem Engagement überhaupt aussagekräftige Daten generiert.[245] Insbesondere Bemühungen, den individuellen Lernerfolg auf Grundlage solcher Daten vorherzusagen[246], sind kritisch zu beurteilen. Lernverhalten und -erfolg dürfen nicht auf einfach erfassbare Daten reduziert werden. Einem solchen Reduktionismus liegt eine behavioristische Betrachtungsweise des Menschen zugrunde. Kognitive Verstehensprozesse und konstruktive Lernszenarien werden dabei unter Umständen zurückgedrängt. Wenn es jedoch um die Reichweite von Bildung geht, die auch der Persönlichkeitsentwicklung und der Bildung junger Menschen zu verantwortlichen Personen mit Urteilskraft dienen soll, dann sind die personale Vermittlung und die personalen Aspekte von Bildung nicht zu vernachlässigen.

Datengetriebene, KI-gestützte Lehr- und Lernsysteme können den jeweiligen Lernprozess unterstützen. Die Bildungsvorteile hinsichtlich der Wissens- und Informationsvermittlung durch den Einsatz digitaler Werkzeuge sind nicht zu unterschätzen. Sie ersetzen aber nicht die personale Vermittlung und die personalen Aspekte von Bildung. Das analoge Gespräch ist unverzichtbar für das motivations- und identitätsstiftende Potenzial, das der Unterricht bereitstellen

245 https://www.eetasia.com/orwellian-nonsense-or-innovation-in-the-classroom [12.01.2023].
246 Gray/Perkins 2019.

sollte. Verstehensprozesse, deren Überprüfung ein personales Gegenüber braucht, können von digitalem Lernen allein nicht angestoßen werden. Die Verantwortung für die Lernprozesse liegt bei den Menschen, die die Wahrung und Förderung der Selbstbestimmung auch aufseiten der Lernenden im Blick haben sollten.[247] Wenn Bildung als Beitrag zu einem mündigen Personsein Bestand haben soll, kann sie nicht vollständig automatisiert und an Maschinen delegiert werden, da personale Bildung (stets auch) auf interpersonale und pädagogische Aktion und Interaktion angewiesen ist. Die Relevanz der Schule als Sozialraum der Interaktion zwischen Menschen ist dabei nicht zu unterschätzen. All diese Aspekte in ihrer Bedeutung für den Bildungsbereich hat die Coronapandemie unterstrichen, vor allem die Notwendigkeit des sozialen Miteinanders.[248]

6.6 Fazit und Empfehlungen

In dieser Stellungnahme wird das Schulsystem als ein wesentlicher Ort von Bildung in den Fokus genommen. Ausgeklammert wurden die enormen praktischen Herausforderungen in diesem System, die ganz offenkundig nicht einfach durch den Einsatz von KI-Technologien gelöst werden können, dennoch aber den Kontext der zukünftigen Gestaltung mitbestimmen. Personalmangel, sich verschärfende Lerndefizite bei den Schülerinnen und Schülern – auch verstärkt infolge der Coronapandemie – sowie verschiedene infrastrukturelle Schwächen sind gut bekannt und gleichwohl noch immer nicht adäquat gelöst. Sie bilden den Hintergrund für jede Debatte, ob und

247 Weiterführende Fragen stellen sich in diesem Feld, die vor allem auf der Makroebene angesiedelt sind: Wer ist der Industriepartner oder übernimmt dies die öffentliche Hand, vor allem da es sich um kostenintensive Entwicklungen und teilweise auch Anwendungen handelt? Sind vor allem bei dieser Entwicklung auch Lernende mit besonderen Bedürfnissen im Blick?

248 Deutscher Ethikrat 2022b.

gegebenenfalls wie sich KI-Technologien in ethisch verantwortlicher Weise in das deutsche Schulwesen integrieren lassen.

Übergreifend lässt sich diese Situation dabei sowohl für als auch gegen einen verstärkten Einsatz solcher neuen Systeme ins Feld führen: dafür, weil durch solche Technologien gegebenenfalls einige Defizite und Mängel zumindest partiell gemildert bzw. aufgefangen werden könnten. Ebenso lässt sich angesichts des teils schlechten Zustandes des Schulwesens in Deutschland auch gegen den Einsatz argumentieren, vor allem wenn dieser dazu dient, das grundsätzliche Ziel, Bildung personell, finanziell und strukturell besser zu gestalten, zu umgehen bzw. von der Agenda zu nehmen. Entsprechend gilt es zu betonen, dass die infrastrukturellen und personellen Herausforderungen des Schulsystems weiter – und schon sehr lange – ihrer Verbesserung harren und prioritär angegangen werden sollten; ein Thema, das weit über den Fokus dieser Stellungahme hinausreicht. Zugleich sollten technologische Entwicklungen realistisch und mit Blick darauf untersucht und erwogen werden, ob und gegebenenfalls inwieweit sie auch in dieser Hinsicht Abhilfe schaffen könnten. Potenziale sollten nicht deswegen verschenkt werden, weil sie eventuell nur symptomatische, aber keine tiefgreifenden Lösungen bieten. Einfache Ersatzoptionen oder leichte Lösungen der strukturellen Herausforderungen sind wiederum von den hier diskutierten Technologien nicht zu erwarten und sollten nicht zum Nachlassen eines Reformdrucks im Schulsystem führen.

Eingedenk des hier verwendeten Verständnisses des Bildungsbegriffs sowie der Analyse zur Erweiterung und zum Vermindern menschlicher Fähigkeiten durch Delegation von vormals menschlichen Tätigkeiten an Technologien sollte im schulischen Kontext zudem unterstrichen werden, dass die grundlegende pädagogische Arbeit von den vorgestellten Systemen nicht ersetzt werden kann und sollte. Die Relevanz der Schule als Sozialraum der Interaktion zwischen Lehrenden und Lernenden sollte unbedingt erhalten bleiben. Allerdings

bieten sich durchaus Potenziale, diese Interaktionen zu erweitern. Das Hauptziel beim Einsatz von KI-Systemen im schulischen Unterricht sollte immer die Förderung der Lernenden auf der einen Seite und die Unterstützung wie auch Entlastung der Lehrenden auf der anderen Seite sein.

Wie in anderen Anwendungsbereichen auch lässt sich also nicht pauschal beantworten, welche der vorgestellten Tools – die von Vokabeltrainern und intelligenten Tutorsystemen über Classroom Analytics bis hin zu Telepräsenzrobotern reichen – wann, von wem und in welchem Fach eingesetzt werden sollten. Gesellschaftliche Bildungsvorstellungen, die sich in den Institutionen und ihren Lernzielen niederschlagen, unterschiedliche Verantwortlichkeiten, aber auch die (in der Aus- und Weiterbildung erworbene) Kompetenz der Lehrkräfte, die Methoden gemäß den Lernzielen festzulegen und die digitalen Techniken in ihrer Ambivalenz wahrzunehmen, sowie die Bedarfe aufseiten der Lernenden – all dies sind wichtige Einflussfaktoren, die auf die Notwendigkeit einer einzelfallbezogenen bzw. auf einzelne Anwendungsbereiche bezogenen Abwägung verweisen.

Da das hier angelegte Verständnis der Bildung des Menschen davon ausgeht, dass diese nicht nur in optimierbarer und berechenbarer Anhäufung von Wissen, sondern vor allem in einem konstruktiven und verantwortlichen Umgang mit erlerntem Wissen besteht, ist besondere Aufmerksamkeit geboten, die Lernprozesse, die zentral für die Persönlichkeitsbildung des Menschen sind, nicht auszulagern. Entscheidungen über das Setting des genauen Einsatzes des jeweiligen Tools sollten daher letztendlich immer beim Menschen liegen. Zudem muss eine Beeinträchtigung der Privatsphäre der Lernenden (und der Lehrenden) ausgeschlossen werden.

Empfehlungen

›› *Empfehlung Bildung 1:* Digitalisierung ist kein Selbstzweck. Der Einsatz sollte nicht von technologischen Visionen, sondern von grundlegenden Vorstellungen von Bildung, die

auch die Bildung der Persönlichkeit umfassen, geleitet sein. Die vorgestellten Tools sollten deshalb im Bildungsprozess kontrolliert und als ein Element innerhalb der Beziehung zwischen Lehrenden und Lernenden eingesetzt werden.

» *Empfehlung Bildung 2:* Für jedes Einsatzgebiet gilt es, eine angemessene Abwägung von Chancen und Risiken vorzunehmen. Insbesondere sollten Autonomie und Privatheit von Lehrenden und Lernenden hohen Schutz erfahren. Besondere Chancen ergeben sich im Bereich der Inklusion und Teilhabe, wo das Potenzial dieser Systeme genutzt werden sollte, um etwa sprachliche oder räumliche Barrieren abzubauen.

» *Empfehlung Bildung 3:* Tools, die einzelne Elemente des Lehr- und Lernprozesses ersetzen bzw. ergänzen (enge Ersetzung) und nachweislich Fähigkeiten, Kompetenzen oder soziale Interaktion der Personen, die sie nutzen, erweitern, wie etwa einige intelligente Tutorsysteme oder Telepräsenzroboter für externe Lehrbeteiligung, sind prinzipiell weniger problematisch als solche, die umfassendere bzw. weitere Teile des Bildungsprozesses ersetzen. Je höher der Ersetzungsgrad, desto strenger müssen Einsatzbereiche, Umgebungsfaktoren und Nutzenpotenziale evaluiert werden.

» *Empfehlung Bildung 4:* Es gilt, standardisierte Zertifizierungssysteme zu entwickeln, die anhand transparenter Kriterien des Gelingens von Lernprozessen im genannten umfassenden Sinne Schulämter, Schulen und Lehrkräfte dabei unterstützen können, sich für oder gegen die Nutzung eines Produkts zu entscheiden. Hier kann sich auch der Empfehlung zur dauerhaften Einrichtung länderübergreifender Zentren für digitale Bildung, wie es im Gutachten „Digitalisierung im Bildungssystem: Handlungsempfehlungen von der Kita bis zur Hochschule" der Ständigen Wissenschaftlichen Kommission der Kultusministerkonferenz angesprochen wurde, angeschlossen werden.

- *Empfehlung Bildung 5:* Bei der Entwicklung, Erprobung und Zertifizierung entsprechender KI-Produkte bedarf es einer engen Zusammenarbeit mit den relevanten Behörden, mit den jeweils zuständigen pädagogischen Fachgesellschaften sowie der Partizipation von Beteiligten, um Schwachstellen der Produkte frühzeitig zu entdecken und hohe Qualitätsstandards zu etablieren. Bekannte Herausforderungen KI-getriebener Technologien wie beispielsweise Verzerrungen bzw. Bias oder Anthropomorphisierungstendenzen sollten bei der Entwicklung und Standardisierung berücksichtigt werden.
- *Empfehlung Bildung 6:* Um den verantwortlichen Einsatz von KI-Technologien im Bildungsprozess zu gewährleisten, muss die Nutzungskompetenz insbesondere der Lehrkräfte erhöht werden; es bedarf der Entwicklung und Etablierung entsprechender Module und Curricula in der Aus-, Fort- und Weiterbildung. Insbesondere die Gefahren eines verengten pädagogischen Ansatzes und eines Deskillings in der Lehre sollten dabei aktiv in den Blick genommen werden. Ebenso sollte die digitale Nutzungskompetenz von Lernenden sowie Eltern gestärkt und um KI-Aspekte erweitert werden.
- *Empfehlung Bildung 7:* Im Sinne der Beteiligungsgerechtigkeit sollten KI-basierte Tools Lernenden grundsätzlich auch für das Eigenstudium zur Verfügung stehen.
- *Empfehlung Bildung 8:* Die Einführung von KI-Tools im Bildungsbereich erfordert ferner den Ausbau verschiedener flankierender Forschungsbereiche. Sowohl theoretische Fundierung als auch empirische Evidenz zu Effekten, etwa auf die Kompetenzentwicklung (z. B. Problemlösen), oder zur Beeinflussung der Persönlichkeitsentwicklung bei Kindern und Heranwachsenden muss weiter ausgebaut werden. Dabei sollte nicht nur stärker in Forschung und entsprechende Produktentwicklung investiert, sondern vor allen Dingen auch die praktische Erprobung und Evaluation im schulischen Alltag verstärkt werden.

» *Empfehlung Bildung 9:* Des Weiteren stellt sich hier die Problematik der Datensouveränität. Zum einen sind bei der Sammlung, Verarbeitung und Weitergabe von bildungsbezogenen Daten strenge Anforderungen an den Schutz der Privatsphäre zu beachten. Zum anderen sollte die gemeinwohlorientierte, verantwortliche Sammlung und Nutzung von großen Daten, etwa in der prognostischen lehrunterstützenden Anwendung, ermöglicht werden.

» *Empfehlung Bildung 10:* Eine vollständige Ersetzung von Lehrkräften läuft dem hier skizzierten Verständnis von Bildung zuwider und ist auch nicht dadurch zu rechtfertigen, dass schon heute in bestimmten Bereichen ein akuter Personalmangel und eine schlechte (Aus-)Bildungssituation herrschen. In der komplexen Situation der schulischen Bildung bedarf es eines personalen Gegenübers, das mithilfe technischer Komponenten zwar immer stärker unterstützt werden kann, dadurch selbst als Verantwortungsträger für die pädagogische Begleitung und Evaluation des Bildungsprozesses aber nicht überflüssig wird.

» *Empfehlung Bildung 11:* In Anbetracht der erkenntnistheoretischen und ethischen Herausforderungen und unter Abwägung potenzieller Nutzen und Schäden stehen die Mitglieder des Deutschen Ethikrates dem Einsatz von Audio- und Videomonitoring im Klassenzimmer insgesamt skeptisch gegenüber. Insbesondere erscheint die Analyse von Aufmerksamkeit und Emotionen per Audio- und Videoüberwachung des Klassenraums mittels aktuell verfügbarer Technologien nicht vertretbar. Ein Teil des Ethikrates schließt den Einsatz von Technologien zur Aufmerksamkeits- und Affekterkennung zukünftig jedoch nicht vollständig aus, sofern sichergestellt ist, dass die erfassten Daten eine wissenschaftlich nachweisbare Verbesserung des Lernprozesses bieten und das hierfür notwendige Monitoring von Lernenden und Lehrkräften keine inakzeptablen Auswirkungen auf deren Privatsphäre und Autonomie hat. Ein anderer Teil des Ethikrates hingegen hält die

Auswirkungen auf Privatsphäre, Autonomie und Gerechtigkeit hingegen generell für nicht akzeptabel und befürwortet daher ein Verbot von Technologien zu Aufmerksamkeitsmonitoring und Affekterkennung in Schulen.

7 ÖFFENTLICHE KOMMUNIKATION UND MEINUNGSBILDUNG

7.1 Einleitung

Durch die digitale Transformation und die in der Folge veränderten politisch relevanten Kommunikationsprozesse wird die Demokratie als Herrschafts- und als Lebensform verändert.[249] Die rasante Verbreitung digitaler Plattformen und sozialer Medien mit ihren algorithmisch vermittelten Informations- und Kommunikationsangeboten wirkt sich nicht nur auf einzelne gesellschaftliche Sphären aus, sondern potenziell auch auf große Teile der öffentlichen Kommunikation und Meinungsbildung – mit Konsequenzen für das demokratische Legitimationsgefüge, im Positiven wie im Negativen.

Über die seit der Pionierzeit der Digitalisierung bestehenden hohen Erwartungen, durch mehr Beteiligung, weitreichende Transparenz, bessere Kommunikation und schnellere Bewältigung hoheitlicher Aufgaben die Verwandlung demokratischer Strukturen zum Besseren oder gar demokratische Umbrüche erst anzustoßen oder zu ermöglichen, hat sich ein Schatten gelegt: die immer deutlicher zutage tretende systemische Verletzlichkeit politischer Institutionen durch Desinformations- und Manipulationsversuche und Polarisierungstendenzen. Was zu Beginn durch mehr Partizipation, erhebliche Vergrößerungen der Zugänge zu Informationen und Abbau von Hierarchien vor allem als Ausweitung und Vertiefung der Möglichkeiten eines demokratischen Gemeinwesens und seiner miteinander agierenden Bürgerinnen und Bürger angelegt zu sein schien, erweist sich in Teilen schon jetzt als gefährdend. Zunehmende Diskursverrohung und Vertrauenserosion können dazu führen, dass sich die realen Freiräume zur

249 Nationale Akademie der Wissenschaften Leopoldina 2021.

Partizipation und zum gemeinsamen Ringen um das bessere Argument eher verringern.

Nimmt man die Chancen wie Risiken zusammen, erscheint die Algorithmisierung politischer Kommunikationsprozesse oder gar demokratisch legitimierter Verfahren als gestaltungsbedürftig. Dabei sind die vielfach als größte Bedrohung wahrgenommenen Gefährdungen nicht die Möglichkeiten politisch motivierter Manipulation oder Angriffe auf digitale Infrastrukturen. Die Herausforderung liegt tiefer. Sie berührt das Selbstverständnis mündiger, in ihren politischen Entscheidungen grundsätzlich freier Personen. Es geht also um die Gefahr einer Verminderung menschlicher Autorschaft. Vor allem die Informations- und die Diskursqualität, die für die je eigene Willensbildung eminent wichtig sind, werden von den sozialen Medien herausgefordert.

7.2 Das Internet als soziotechnisches System: Funktionsweise sozialer Medien

7.2.1 Neue soziotechnische Infrastrukturen

Die Möglichkeiten zur Informationsbeschaffung und Kommunikation im Internet haben sich in den vergangenen Jahrzehnten drastisch verändert, insbesondere durch die rasch angestiegene Verbreitung und Marktmacht interaktiver Plattformen, über die inzwischen große Teile des digitalen Informationsaustauschs ablaufen. Hierzu gehören nicht nur soziale Netzwerke wie Facebook, Instagram, Twitter, YouTube, TikTok oder LinkedIn, sondern auch Suchmaschinen, allen voran Google, und Messengerdienste zum Austausch von Sofortnachrichten wie WhatsApp, Signal, Telegram oder WeChat.

Die Übergänge zwischen sozialen Netzwerkplattformen, Suchmaschinen und Messengerdiensten entwickeln sich dabei fließend und es kommt auch bei inhaltlichen und funktionalen

Merkmalen zunehmend zu Konvergenzen. Viele Plattformen bieten inzwischen einander ähnelnde Möglichkeiten dahingehend an, multimediale Inhalte zu erstellen und zu veröffentlichen oder live zu verbreiten, auf die Inhalte anderer zu reagieren, sie zu bewerten, zu kommentieren und – auch plattformübergreifend – weiterzuverbreiten, sich über direkte Nachrichten und Konferenzschaltungen mit anderen Personen auf der Plattform auszutauschen, die Plattform nach Inhalten, Profilen, Gruppen oder sonstigen Angeboten zu durchsuchen und bestimmte Inhalte zu abonnieren. Auch Optionen, eigene Inhalte gezielt zu bewerben und Produkte und Dienstleistungen direkt anzubieten oder zu kaufen, sind vielfach vorhanden.

Fast alle weiter verbreiteten Plattformen und Dienste werden von privaten Unternehmen aus den USA oder China betrieben und die größten sozialen Netzwerke gehören nur wenigen Firmen. Von den zehn am meisten genutzten Angeboten werden vier – Facebook, WhatsApp, Instagram und Messenger – vom US-Konzern Meta betrieben. Das zweitgrößte soziale Netzwerk, die Videoplattform YouTube, gehört zum US-Konzern Google, der gleichzeitig mit seiner Web-Suchmaschine und Onlinediensten wie Google Fotos, Gmail, Google Maps, Google Cloud und Google Drive große Teile des Internets dominiert. Drei weitere der zehn meistgenutzten Dienste Angebote werden von den chinesischen Unternehmen Tencent (WeChat) und ByteDance (TikTok und seine chinesische Version Douyin) angeboten.[250]

Aufgrund dieser Marktmacht sowie der Vielseitigkeit und Integration der von einzelnen Konzernen bereitgehaltenen Dienste funktionieren die Angebote der großen Konzerne inzwischen als reichhaltige soziotechnische Infrastrukturen, in denen sich ein Großteil des Online-Nutzungsverhaltens nach den Vorgaben weniger Konzerne abspielt. Zahlreiche Einbindungsmöglichkeiten für externe Webseiten und Dienste

250 Kemp 2023, Abb. 182.

verstärken diesen Effekt noch, wie beispielsweise das Teilen von Inhalten anderer Webseiten auf der Plattform, die direkte Verknüpfung externer Dienste mit einer Plattform und die Option, sich per Einmalanmeldung (Single Sign-on) mit den Anmeldedaten großer Portale wie Google oder Facebook auch bei anderen Seiten zu authentifizieren. Wie weitumfassend einzelne Plattformen die Interneterfahrung prägen können, zeigt sich beispielsweise an Facebooks Initiative internet.org, die seit 2014 in ausgewählten Ländern über Mobilgeräte kostenlosen Zugang zum Internet anbietet, aber die Registrierung bei Facebook voraussetzt. Ebenso ist die chinesische App WeChat aufgrund ihres großen Funktionsumfangs einschließlich Bezahlfunktion für viele Menschen in China zum Zentrum fast all ihrer Onlineaktivitäten geworden.

7.2.2 Informationsauswahl und Kuratierung

Mit der Fülle der Informationen und Interaktionsmöglichkeiten in sozialen Medien gehen technische Herausforderungen und ökonomische Potenziale einher, die gemeinsam zur Gestaltung aktueller Funktionsweisen und Geschäftsmodelle beigetragen haben. Im Gegensatz zu klassischen Medien, die Inhalte mit einer begrenzten Zahl mitarbeitender Personen nach einem bestimmten Konzept erstellen, auswählen, prüfen und veröffentlichen, produzieren in sozialen Medien neben professionellen Akteuren auch die Nutzerinnen und Nutzer selbst die dort zirkulierenden Beiträge. So entstehen schnell riesige Mengen an Inhalten von höchst unterschiedlicher Qualität. Dies stellt die Plattformen wie auch ihre Kundschaft vor das Problem der Informationsauswahl, das sich mit traditionellen redaktionellen Strukturen, in denen Profis Inhalte aussuchen und qualitativ prüfen, nicht mehr bewältigen lässt. Aktuelle Ansätze zur Lösung dieses Problems delegieren daher einen Großteil der redaktionellen Arbeit, vor allem die Kuratierung, das heißt die Auswahl von Inhalten, an maschinelle Systeme.

Insbesondere die personalisierte Kuratierung, die dazu führt, dass jeder Person beim Besuch einer Plattform auf sie persönlich zugeschnittene Inhalte in einer bestimmten Reihenfolge angezeigt werden, erfolgt inzwischen weitgehend durch Algorithmen, die menschliche Redaktionsarbeit auf dieser Ebene somit nahezu vollständig ersetzen und unsichtbar im Hintergrund arbeiten.

Die Kriterien, nach denen solche Algorithmen ihre Auswahl treffen, sind eng mit ökonomischen Faktoren verknüpft. Die meisten Plattformen und Dienste folgen einem werbebasierten Geschäftsmodell, bei dem Werbetreibende dafür bezahlen, möglichst viele Menschen zu erreichen, die an ihren Produkten interessiert sein könnten. Dies gelingt am besten, wenn die Interessen der einzelnen Nutzerinnen und Nutzer erstens möglichst präzise bekannt sind und Menschen zweitens möglichst viel Zeit auf der Plattform verbringen. Je mehr Zeit und Aufmerksamkeit eine Person der Plattform widmet, desto besser sind die Voraussetzungen, ihr auf persönliche Interessen zugeschnittene Werbung zu präsentieren. Dies geschieht auf Basis der Annahme, dass derart maßgeschneiderte Werbung größere Erfolgsaussichten hat als Werbung, die weniger eng orientiert an der Aufmerksamkeit und den Interessen ihrer Zielgruppe präsentiert wird.

Vor diesem Hintergrund lohnt es sich für Plattformen, so viele Datenspuren wie möglich über den persönlichen Hintergrund, die Interessen, das Nutzungsverhalten und das soziale Netzwerk der Personen, die sie nutzen, zu sammeln und diese Daten dazu zu verwenden, deren Aufmerksamkeit mithilfe personalisierter Inhalte möglichst lange zu fesseln und dabei auf ebenfalls personalisierte Werbeinhalte zu lenken. Dies gelingt technisch mithilfe von Empfehlungsdiensten (Collaborative Filtering Recommender Systems), das heißt Systemen, die neben den eigenen Informationen und nutzungsbasierten Vorlieben jeder Person zusätzlich die entsprechenden Daten von Menschen aus deren Netzwerk oder mit ähnlichem Profil sowie die jeweiligen Interaktionen berücksichtigen

(siehe Infokasten 7). Auf Grundlage dieser wechselseitigen Interessenanalyse können solche Algorithmen mit hoher und im Zuge der Nutzungsdauer ständig wachsender Präzision vorhersagen, welche Inhalte jemand angeboten bekommen sollte, um das weitere Verweilen dieser Person auf der Plattform zu maximieren. Weitere Mechanismen, die von Plattformen eingesetzt werden, um die Zeit, die Menschen auf eine Plattform verbringen, auszudehnen, sind das ständige Nachladen neuer Inhalte am Ende einer Seite (Infinite Scroll) und der Einsatz von Benachrichtigungen, die auf neue abonnierte Inhalte, Kommentare oder sonstige Reaktionen hinweisen und so dazu motivieren, die Plattform erneut aufzurufen.

Die Übergänge zwischen Werbung und regulären Inhalten und Interaktion (sogenannte organische Inhalte) sind oft fließend. Viele Firmen, Organisationen und sonstige Anbieter, die Werbung auf Plattformen schalten, betreiben gleichzeitig Profile, auf denen sie auch reguläre Beiträge veröffentlichen und mit anderen im Netzwerk interagieren. Reguläre Beiträge können auch nachträglich gegen Geld eine größere Reichweite erhalten und erscheinen dann ebenfalls als Anzeige. Hinzu kommt, dass Privatpersonen, die mit erfolgreichen Inhalten auf einer Plattform eine große Zahl an Profilen, die ihre Neuigkeiten abonnieren (Follower), und somit viel Reichweite erlangt haben, zunehmend in Kooperationen mit Werbetreibenden als Influencer agieren, die Produkte und Dienstleistungen gegenüber ihrem Publikum gezielt rezensieren und bewerben.

Infokasten 7: Facebook-Algorithmus

Die Auswahlkriterien von Empfehlungsalgorithmen auf sozialen Plattformen werden in der Regel nicht vollumfänglich veröffentlicht, wohl jedoch bestimmte Akzentuierungen und Änderungen, wie folgendes Beispiel zeigt.

Nachdem Facebook 2006 in der Ursprungsversion ihres Newsfeeds nur Inhalte aus unmittelbar vernetzten Profilen in chronologischer Reihenfolge angezeigt hatte, wurde mit der Einführung des „Gefällt-mir"-Buttons

ab 2009 erstmals die Möglichkeit angeboten, Inhalte zu bewerten.[251] Diese Interaktionen wurden fortan dafür genutzt, Beiträge mit vielen Likes im Newsfeed nach oben zu spülen. Im Laufe der folgenden Jahre kamen diverse weitere Kriterien dazu, darunter:

- höhere Priorität für Beiträge aus Profilen, mit dem man viel interagiert (2013)[252]
- weniger Reichweite für Profile, die in ihren nicht als Anzeigen geschalteten Beiträgen Werbung verbreiten (2015)[253]
- weniger Reichweite für Beiträge, die von vielen als unwahr markiert werden (2015)[254]
- höhere Priorität für Beiträge, die Ähnlichkeit mit Inhalten haben, mit denen jemand zuvor viel Zeit verbracht hat (2016)[255]
- höhere Priorität für Beiträge, die Ähnlichkeit mit Inhalten haben, auf die jemand nicht mit dem klassischen „Gefällt mir" reagiert hat, sondern mit einem der 2016 eingeführten Emojis „Liebe", „Haha", „Wow", „Traurig" und „Wütend" (2017)[256]
- mehr Reichweite für Beiträge, die besonders häufig geteilt werden und viele Reaktionen und Diskussionen in den Kommentaren auslösen (2018)[257]

Laut Angaben von Facebook[258] sind die „drei hauptsächlichen Signale", die der Algorithmus aktuell bei der personalisierten Priorisierung von Inhalten verwendet:

(1) wie häufig eine Person mit der Quelle, aus der ein Beitrag stammt, zuvor interagiert hat
(2) welche Art von Inhalten eine Person zuvor am häufigsten genutzt hat (z. B. Fotos, Videos oder Links)
(3) wie viele Interaktionen ein Beitrag erhalten hat, insbesondere von anderen im persönlichen Netzwerk

Zusätzlich zu diesen algorithmisch vorgenommenen Gewichtungen können Nutzerinnen und Nutzer einige Kriterien auch selbst festlegen, wie beispielsweise eine Favoritenliste der Profile und Seiten, denen sie folgen und deren Inhalte sie prioritär sehen möchten.

---

251 http://web.archive.org/web/20090212011228/http://blog.facebook.com/blog.php?post=53024537130 [12.01.2023].
252 https://www.facebook.com/business/news/News-Feed-FYI-A-Window-Into-News-Feed [12.01.2023]; https://martech.org/facebook-story-bumping-last-actor [12.01.2023].
253 https://about.fb.com/news/2014/11/news-feed-fyi-reducing-overly-promotional-page-posts-in-news-feed [12.01.2023].
254 https://about.fb.com/news/2015/01/news-feed-fyi-showing-fewer-hoaxes [12.01.2023].
255 https://about.fb.com/news/2016/08/news-feed-fyi-showing-you-more-personally-informative-stories [12.01.2023].
256 https://about.fb.com/news/2016/02/reactions-now-available-globally [12.01.2023]; https://mashable.com/article/facebook-reactions-news-feed [12.01.2023].
257 https://about.fb.com/news/2018/01/news-feed-fyi-bringing-people-closer-together [12.01.2023].
258 https://www.facebook.com/formedia/tools/feed [12.01.2023].

Insgesamt kommt es so auf Plattformen und in sozialen Netzwerken zu einer ausgeprägten Verquickung technisch wie menschlich vermittelter Stimuli und Strukturen, die große soziale und psychologische Wirkmacht entfalten und gemeinsam dazu beitragen, Aufmerksamkeit zu erregen und zu fesseln, die Verweildauer von Menschen auf der Plattform zu maximieren und so letztlich die Rezeption von Werbehinhalten zu optimieren.

7.2.3 Moderation von Inhalten

Eine algorithmisch gesteuerte personalisierte Informationsauswahl, in der ökonomische und aufmerksamkeitsbasierte Faktoren derart eng verbunden sind und die sich anhand des Nutzungsverhaltens ständig weiterentwickelt, führt dazu, dass Inhalte, die besonders sensationell erscheinen oder intensive emotionale Reaktionen auslösen, sich überproportional schnell und weit verbreiten. Dies begünstigt unter anderem Falschnachrichten[259] und Beiträge, auf die mit Wut reagiert wird[260], so zum Beispiel Inhalte wie Hassreden, Beleidigungen und Volksverhetzungen. Das stellt Plattformen vor die Herausforderung, wie solche potenziell problematischen und gleichzeitig verbreitungsstarken Inhalte erkannt werden können und wie mit ihnen umzugehen ist.

In Reaktion auf diese Herausforderung bemühen sich Plattformen darum, ihre Inhalte nach verschiedenen Kriterien zu moderieren (Content-Moderation). Hierbei sind einerseits Menschen beteiligt – sowohl Angestellte der Firmen selbst als auch Personen, die für externe Dienstleister arbeiten, und die Menschen, die das Plattformangebot nutzen und Inhalte melden, die ihnen unangenehm auffallen. Andererseits kommen auch algorithmische Systeme zum Einsatz. In diesem

259 Vosoughi/Roy/Aral 2018.
260 Fan et al. 2014.

soziotechnischen Zusammenspiel geht es zum einen um die Gestaltung und Umsetzung von Regeln, die bestimmen, welche Inhalte erlaubt sind und welche gelöscht werden müssen, zum anderen aber auch darum, welche Inhalte im Rahmen einer algorithmischen Kuratierung gegebenenfalls lediglich in ihrer Verbreitung und Sichtbarkeit eingeschränkt werden.

Grundlage für die Moderation sind zunächst rechtliche Vorgaben, die bestimmte Inhalte als unzulässig einstufen, vor allem solche des Strafrechts; aber auch zivilrechtlich geschützte Persönlichkeitsrechte setzen der Kommunikation Grenzen. Auch die Anforderungen für die Meldung, Prüfung und Entfernung solcher Inhalte wie auch für Beschwerden gegen solche Eingriffe ergeben sich zunächst aus den allgemeinen rechtlichen Regelungen für rechtswidrige Äußerungen. Im deutschen Kontext ist darüber hinaus insbesondere auf das Gesetz zur Verbesserung der Rechtsdurchsetzung in sozialen Netzwerken (Netzwerkdurchsetzungsgesetz) hinzuweisen, welches Plattformbetreiber verpflichtet, offensichtlich illegale Inhalte innerhalb von 24 Stunden nach Eingang der Beschwerde zu löschen oder den Zugang zu ihnen zu sperren (§ 3 Abs. 2 Nr. 2). Das Gesetz selbst definiert nicht, was illegal ist, sondern verweist auf strafrechtliche Normen. An das Konzept des Netzwerkdurchsetzungsgesetzes knüpft der 2022 von der Europäischen Union erlassene Digital Services Act an, der das Netzwerkdurchsetzungsgesetz wohl jedenfalls teilweise ersetzen wird.[261]

Hinzu kommen in der Regel Kommunikationsregeln, die von Plattformbetreibern selbst entwickelt werden (Community-Standards), auf deren Grundlage auch rechtlich zulässige Inhalte gelöscht, gesperrt oder in ihrer Reichweite eingeschränkt werden können. Hierzu gibt es eine kontroverse Diskussion, ob und wann solche Praktiken die Meinungsfreiheit auf unzulässige Weise beschränken.[262]

261 Weiden 2022.
262 Raue 2018.

Moderationsvorgänge können sowohl durch die Meldung anstößiger Inhalte, die Einzelpersonen bei der Nutzung des Plattformangebots auffallen, als auch durch Algorithmen, die beispielsweise nach bestimmten Schlüsselwörtern suchen, ausgelöst werden. Bei Inhalten, die als potenziell problematisch markiert werden, muss anschließend geprüft werden, ob sie die aktuell geltenden Regeln verletzen. Auch an diesen Prüfaufgaben sind Menschen und Maschinen gemeinsam beteiligt. Menschliche Moderation erfolgt typischerweise durch Personen, die häufig unter äußerst prekären Arbeitsbedingungen bei Drittanbietern angestellt sind, mit denen eine Plattform vertraglich zusammenarbeitet.[263] Sie sichten gemeldete Inhalte und gleichen sie mit den jeweils geltenden Regeln ab. Dabei werden sie häufig mit extrem belastendem Material wie Tötungen, Kindesmissbrauch, Tierquälerei und Suizid konfrontiert. Zudem müssen sie innerhalb weniger Sekunden sprachlich und kulturell komplexe Nuancen berücksichtigen, von denen die Zulässigkeit eines Beitrags entscheidend abhängen kann. So kann es zum Beispiel einen großen Unterschied machen, ob ein Beitrag als Satire gewertet wird oder nicht oder ob eine anstößige Botschaft im Kontext des Beitrags nur zitiert und gegebenenfalls kritisch reflektiert oder sich zu eigen gemacht wird.

Der Einsatz von Algorithmen zur Moderation wirkt vielversprechend, da er zum einen anstößige, beispielsweise brutale Inhalte herausfiltern kann, ohne dass diese von Menschen angesehen werden müssen, und zum anderen mit der unübersichtlichen Menge an Daten und Inhalten im Netz besser umgehen kann. Allerdings sind automatisierte Methoden jedenfalls bislang häufig unzureichend, um den kulturellen und sozialen Zusammenhang einer Äußerung einzubeziehen und diese damit adäquat zu beurteilen, da aufgrund der beschriebenen

263 Siehe dazu etwa den Dokumentarfilm „The Cleaners“ aus dem Jahr 2018 (www.thecleaners-film.de); https://time.com/6253180/meta-kenya-lawsuit-motaung [28.02.2023]. Die gleiche Problematik zeigt sich auch bei ChatGTP (https://time.com/6247678/openai-chatgpt-kenya-workers [06.02.2023]).

komplexen Kontexte und Nuancen sowie durch absichtlich verwendete Stilmittel wie Ironie und Sarkasmus Unschärfen entstehen, die sich algorithmisch noch nicht immer zuverlässig zuordnen lassen.[264] Genau solche kontextuellen Faktoren spielen bei der juristischen sowie moralischen Beurteilung von Äußerungen jedoch eine sehr große Rolle.

So kommt es zum Beispiel darauf an, wer gegenüber wem in welchem Kontext einen Begriff verwendet, um zu beurteilen, ob es sich hier um eine (rassistische) Beleidigung handelt. Im amerikanischen Raum konnte etwa gezeigt werden, dass der Einsatz automatisierter Content-Moderation-Techniken verschiedene gesellschaftliche Gruppen unterschiedlich betrifft, da bestimmte Sprachformen beispielsweise eher markiert werden. Facebooks Ansatz, allgemeingültige und „blinde" Regeln zu erschaffen, führte hier dazu, dass Gruppen wie „white men" eher vor Hassrede geschützt wurden als vulnerablere Gruppen wie „black children".[265]

Eine Sorge bei der Content-Moderation liegt in der Gefahr, dass systematisch auch Inhalte gelöscht oder unzugänglich gemacht werden, die nicht gegen (vom Staat oder von Plattformen gesetzte) Regeln verstoßen (sogenanntes Overblocking). Dass es dazu kommt, ist jedenfalls dann plausibel, wenn der Rechtsrahmen Anreize setzt, Inhalte im Zweifel zu löschen, etwa wenn Sanktionen für den Fall der Nichtlöschung vorgesehen sind, nicht aber für die fehlerhafte Löschung. Dies wird beispielsweise in kritischen Betrachtungen des Netzwerkdurchsetzungsgesetzes von Kritikern angenommen.[266] Ob es tatsächlich zu Overblocking kommt, ist schwer empirisch zu prüfen, nicht nur weil dazu Datenzugang benötigt wird, sondern auch weil eine angemessene Prüfung von

264 Algorithmische Methoden zur Sentiment Analysis, die auch Stilmittel wie zum Beispiel Ironie und Sarkasmus erkennen können, werden ständig weiterentwickelt (Sarsam et al. 2020).

265 https://www.propublica.org/article/facebook-hate-speech-censorship-internal-documents-algorithms [12.01.2023]; Davidson/Bhattacharya/Weber 2019.

266 Überblick bei Liesching et al. 2021; vgl. Fichtner 2022.

Kommunikationsinhalten eine Auslegung der Äußerung und eine Berücksichtigung des Kontextes erfordert, was jedenfalls den Aufwand erhöht.

7.2.4 Auswirkungen auf die Erweiterung und Verminderung menschlicher Handlungsfähigkeiten

Die beschriebenen Funktionsweisen von Plattformen und sozialen Medien sind dadurch gekennzeichnet, dass erstens große Teile der Auswahl und Moderation von Inhalten an Algorithmen delegiert werden, dass es dabei zweitens zu komplexen Wechselwirkungen zwischen diesen technisch vermittelten Prozessen und den in diesen Netzwerken agierenden Menschen kommt und dass hier drittens ganz überwiegend kommerziell motivierte Gestaltungsparameter eingesetzt werden, die dazu dienen, die Aufmerksamkeit und Verweildauer von Menschen auf der jeweiligen Plattform zu maximieren. Mithilfe dieser soziotechnischen Verquickungen können menschliche Handlungsfähigkeiten in unterschiedlicher Weise erweitert oder vermindert werden.

Die Delegation von Kuratierungsprozessen an Algorithmen, die aufgrund datenreicher Profile ein überaus effektives personalisiertes Angebot von Inhalten zusammenstellen, erweitert zunächst die Möglichkeiten von Plattformbetreibern und Werbetreibenden, ein großes Publikum erfolgreich zu erreichen. Für einzelne Nutzerinnen und Nutzer ist die Personalisierung des Angebots mit Komfort- und Effizienzgewinnen verbunden, weil gewünschte beziehungsweise zu den eigenen Interessen passende Inhalte schnell auffindbar sind oder proaktiv angeboten werden. Auch dies kann eine Erweiterung von Handlungsmöglichkeiten bedeuten, zum Beispiel wenn persönliche Ziele so besser oder schneller erreicht werden können oder wenn aufgrund der effektiven Delegation der

Inhaltsauswahl an Algorithmen Entlastungseffekte auftreten, die Freiräume für andere Aktivitäten schaffen.

Die soziotechnisch hybride Moderation von Inhalten erweitert zudem die Möglichkeiten aller Beteiligten, eine sehr große Inhaltsfülle und -vielfalt verbreiten und rezipieren zu können, ohne dabei ein Übermaß problematischen Materials in Kauf nehmen zu müssen. Die beschriebenen, hierbei bestehenden Unzulänglichkeiten und Abwägungsschwierigkeiten können jedoch, wie in Abschnitt 7.4 noch näher beleuchtet wird, auch vermindernd auf die Diskursqualität wirken.

Eine Verminderung menschlicher Handlungsspielräume kann sich ebenso ergeben, wo es Menschen schwerfällt, sich dem Sog von Plattformangeboten zu entziehen und die Nutzung dieser Angebote auf ein für sie gesundes Maß zu beschränken. Die soziotechnische Optimierung von Plattformangeboten zur Maximierung von Aufmerksamkeit und Verweildauer entwickelt hier große Wirkmacht; sogar regelrechte Abhängigkeiten können entstehen.[267] Persönliche Freiheit wird zum Beispiel dann vermindert, wenn das selbstbestimmte Verfolgen eigentlich gesetzter Ziele erschwert wird, weil eine Loslösung vom Plattformangebot nicht rechtzeitig hinreichend gelingt.

Auch die Delegation der Inhaltsauswahl an eine algorithmische Kuratierung kann Autorschaft vermindern, insbesondere wenn eine rationale Auseinandersetzung mit Alternativen aufgrund der algorithmischen Vorwegnahme bestimmter Relevanzentscheidungen nur noch eingeschränkt stattfinden kann. Die auf Grundlage solcher Vorwegnahmen entstehenden personalisierten und soziotechnisch komplexen Strukturen, innerhalb derer Inhalte zur Auswahl angeboten werden (Choice Architectures), können ausgeprägte Nudging-Effekte entfalten, die menschliche Entscheidungen wirkmächtig beeinflussen, ohne dass dies für die Betroffenen erkenntlich ist (vgl. Abschnitt 4.5). Dies ist als besonders problematisch

267 Cheng et al. 2021.

einzustufen, wenn die algorithmischen Auswahlkriterien kommerziell motiviert oder von anderen Interessen privater Konzerne bestimmt werden, die von außen kaum nachvollziehbar sind und deren Ausrichtung am Gemeinwohl nicht vorausgesetzt werden kann.[268]

Neben diesen allgemeinen Auswirkungen der Funktionsweisen von Plattformen und sozialen Medien verändern sich auch die Informations- und die Diskursqualität, die wichtige Grundlagen der öffentlichen Meinungsbildung sind – mit potenziell weitreichenden Konsequenzen für Prozesse der politischen Willensbildung. Hier steht eine Reihe von Phänomenen in der Diskussion, die in den nächsten beiden Abschnitten näher betrachtet werden. Wie weit verbreitet und wirkmächtig diese Effekte sind, lässt sich aktuell zwar noch nicht abschließend beurteilen, auch weil die Datenlage mitunter unklar oder widersprüchlich ist. Manche Fragen lassen sich nur schwer untersuchen, da zum Beispiel Daten zur Plattformnutzung und den dort verwendeten Algorithmen auch für die Forschung nur teilweise oder in einigen Fällen wie beispielsweise bei verschlüsselten Messengerdiensten gar nicht öffentlich zugänglich sind. Ein genauerer Blick auf die postulierten Mechanismen lohnt jedoch schon deswegen, weil die von ihnen berührten Prozesse grundlegend für unsere Demokratie sind.

7.3 Informationsqualität

Mit Blick auf die Informationsqualität ist zunächst auf die vielen positiven Veränderungen zu verweisen, die sich trotz der oben beschriebenen Herausforderungen bei der Auswahl, Kuratierung und Moderation von Plattforminhalten aus dem im Internet verfügbaren Informationsangebot ergeben. Plattformen und soziale Medien erweitern die Möglichkeiten vieler Menschen, sich zu informieren, sich Gehör zu verschaffen und

268 Hildebrandt 2022; Yeung 2017.

sich mit anderen auszutauschen und zu vernetzen, ganz erheblich.[269] Zweifelsohne eröffnet die Quellenvielfalt und -fülle auf vielen Plattformen wie auch im Internet insgesamt im Vergleich mit klassischen Informationskanälen wie Massenmedien, Bibliotheken oder dem direkten persönlichen Austausch wesentlich umfangreichere Möglichkeiten, sich schnell und einfach zu informieren. Dies gilt erst recht für Menschen, die offline keinen guten Zugang zu Informationsangeboten haben, wie zum Beispiel in strukturschwachen oder sich erst entwickelnden Regionen, oder für Menschen, die in ihrer Mobilität eingeschränkt sind.

Es ist wie noch nie zuvor möglich, dass Menschen Informationen und Einschätzungen von unterschiedlichsten Seiten und in hoher Detailgenauigkeit einholen können. Dazu tragen zum Beispiel Projekte wie die gemeinnützige, umfassende und ständig wachsende Online-Enzyklopädie Wikipedia bei. Auch wissenschaftliche Erkenntnisse werden durch Open-Access-Modelle leichter zugänglich. Des Weiteren können sich an speziellen Themen Interessierte auf Plattformen, in Gruppen oder Foren oder über die Verwendung relevanter Markierungen von Beiträgen beispielsweise mit Hashtags zielgerichtet austauschen.[270] Es steht online jeder Person frei, die eigene Informationsnutzung eigenständig zu diversifizieren und beispielsweise gezielt nach Quellen und Stimmen zu suchen, die den eigenen Horizont erweitern. Gemeinsam eröffnen diese Optionen erhebliche Chancen, menschliche Handlungsspielräume, Teilhabemöglichkeiten und Räume für den gemeinsamen öffentlichen Vernunftgebrauch zu erweitern.[271]

Demgegenüber wird vielfach die Sorge geäußert, dass sich insbesondere aus den derzeit gängigen Kriterien der algorithmischen Kuratierung mit ihrer Priorisierung fesselnder

269 Ortiz et al. 2019; Donelan 2016.
270 https://crutchesandspice.com/2022/11/16/with-twitter-crumbling-it-feels-like-the-world-is-collapsing-on-disabled-people [12.01.2023].
271 https://www.theguardian.com/technology/2022/nov/30/twitter-journalism-elon-musk [12.01.2023].

Inhalte auch negative Auswirkungen auf die Informationsqualität ergeben. In der Diskussion stehen hier vor allem die Verbreitung von Falschnachrichten, die Entstehung von Filterblasen und Echokammern und ein Trend zu Inhalten, die negative emotionale und moralische Reaktionen und Interaktionen provozieren und so möglicherweise zu den in Abschnitt 7.4 näher beleuchteten gesellschaftlichen Polarisierungs- oder Fragmentierungstendenzen beitragen.

7.3.1 Falschnachrichten und Verschwörungstheorien

Die bereits erwähnte Tendenz, dass sich unter den aktuellen Kuratierungskriterien vieler Plattformen sensationell wirkende Falschinformationen, insbesondere bei politischen Themen[272], besonders schnell verbreiten, ist mit der Sorge verbunden, dass dies selbst dann nachhaltige Effekte auf die Informationsqualität haben kann, wenn solche Inhalte im Rahmen von Moderationsbemühungen erkannt, entfernt, in ihrer Verbreitung gehemmt oder mit einordnenden Hinweisen versehen werden.

In diesem Zusammenhang wird häufig auch die Annahme vorgebracht, dass populistische Strömungen, Verschwörungstheorien[273] und weitere politisch motivierte Bewegungen, bei denen die Erzeugung und Verstärkung bestimmter Stimmungslagen in der Bevölkerung eine wichtige Rolle spielen, in den vergangenen Jahren zugenommen beziehungsweise an Zulauf und/oder Bedeutung gewonnen haben und dass Falschnachrichten in diesem Rahmen gezielt für politische Zwecke eingesetzt werden.[274] Dieser Eindruck wird unter anderem dadurch begünstigt, dass populistische und

272 Vosoughi/Roy/Aral 2018.
273 Der Begriff der Verschwörungstheorie hat sich eingebürgert, obwohl der Ausdruck „Theorie" angesichts der Fragwürdigkeit bzw. Widersprüchlichkeit der postulierten Inhalte oftmals nicht angemessen ist.
274 Muirhead/Rosenblum 2019.

verschwörerische Elemente bei vielen Ereignissen und Themen Hand in Hand gehen[275] und in jüngerer Zeit eine große Rolle im öffentlichen Diskurs haben.

Es ist nicht von der Hand zu weisen, dass populistische und von verschwörungstheoretischen Ideen geprägte Inhalte demokratiegefährdend wirken können, da sie häufig ausdrücklich darauf abzielen, Vertrauen in demokratisch legitimierte Prozesse und Institutionen zu untergraben, schuldige Personen ausfindig zu machen, die Welt in Gut und Böse aufzuteilen und vermeintliche Gegenseiten zu identifizieren und zu diskreditieren (vgl. Abschnitt 7.4.1).[276] So gibt es beispielsweise empirische Hinweise darauf, dass Menschen, die mit Verschwörungstheorien konfrontiert werden, danach weniger Vertrauen in etablierte Fakten haben und auch weniger Hilfsbereitschaft zeigen.[277]

Ob der Glaube an Verschwörungstheorien in den letzten Jahren tatsächlich zugenommen hat, ist umstritten, auch weil der Definitionsgegenstand oft schwer zu fassen ist und eindeutige empirische Belege fehlen.[278] Die These, dass die in Abschnitt 7.2 beschriebenen Selektionskriterien auf Plattformen auch zur einfachen und schnellen Verbreitung von Verschwörungstheorien und populistischen Inhalten beitragen, erscheint jedoch zumindest plausibel. Falschnachrichten verbreiten sich beispielsweise unter anderem dann besonders erfolgreich, wenn sie Empörung und Wut schüren.[279] Es lässt sich zudem nachweisen, dass Menschen, die soziale Medien intensiv als Informationsquelle nutzen, empfänglicher für konspiratives Gedankengut sind.[280] Zudem lässt die wiederholte Konfrontation mit Falschnachrichten selbige bereits als

275 Castanho Silva/Vegetti/Littvay 2017.
276 https://www.deutschland.de/de/topic/politik/verschwoerungstheorien-und-demokratie-experte-michael-butter [12.01.2023].
277 van der Linden 2015.
278 Uscinski et al. 2022; siehe auch https://www.sciencemediacenter.de/alle-angebote/research-in-context/details/news/glaube-an-verschwoerungstheorien-nahm-im-laufe-der-zeit-nicht-zu [18.01.2023].
279 Chuai/Zhao 2022.
280 Ender et al. 2023.

glaubwürdig erscheinen[281] und besonders einprägsam wirken[282], selbst wenn sie später diskreditiert oder als umstritten markiert werden. Dies gilt erst recht, wenn Inhalte innerhalb eines persönlichen Netzwerks verbreitet werden, dessen Quellen aufgrund gemeinsamer Interessen und Überzeugungen von vornherein als besonders vertrauenswürdig gelten (vgl. Abschnitt 7.3.2).[283]

Moderationsversuche können bei der Eindämmung von Falschnachrichten oft kaum mithalten. Inhalte, die aus vertrauten oder als wirkmächtig wahrgenommenen Quellen stammen oder bereits weit verbreitet wurden und daher als besonders beliebt erscheinen, werden unabhängig vom Wahrheitsgehalt vorzugsweise konsumiert[284] und geteilt[285]. Die gezielte und massenhafte Verbreitung von Falschinformationen über eigens hierfür eingerichtete gefälschte und teilweise automatisch betriebene Konten (Fake-Accounts, Bots[286]) sorgt derweil dafür, dass fragwürdige Inhalte schnell kritische Verbreitungsschwellen erreichen. Beispielsweise werden einer Schätzung zufolge auf Twitter zwei Drittel aller Verweise auf beliebte Webseiten durch Bots verbreitet.[287]

7.3.2 Filterblasen und Echokammern

Filterblasen und Echokammern sind Metaphern, die auf der Hypothese gründen, dass der starke Fokus der algorithmischen Kuratierung auf individuelle Präferenzen dazu führen kann, dass die individuell wahrnehmbare Informationsauswahl

281 Pennycook/Cannon/Rand 2018.
282 Murphy et al. 2019.
283 Del Vicario et al. 2016.
284 Für Jugendliche in den USA vgl. etwa Robb 2020.
285 Allem 2020.
286 Unter einem Bot (von engl. *robot* = Roboter) versteht man ein Computerprogramm, das weitgehend automatisch sich wiederholende Aufgaben abarbeitet, ohne dabei auf eine Interaktion mit einem Menschen angewiesen zu sein.
287 Wojcik et al. 2018.

zunehmend homogener wird. Das Phänomen der Filterblase beschreibt eine potenzielle Verarmung des Themen- und Meinungsspektrums im personalisierten Angebot.[288] Dahinter steht die Vermutung, dass Empfehlungssoftware, sobald sie persönliche Interessen und Vorlieben erkannt hat, vornehmlich Inhalte präsentiert, die dem Material ähneln, auf das man zuvor bereits mit erhöhter Aufmerksamkeit oder Interaktionen reagiert hat, während andere Inhalte, die weniger offensichtlich in das ermittelte Profil passen, immer stärker ausgeschlossen werden. Eine Untersuchung auf Facebook zeigte beispielsweise, dass von persönlichen Kontakten geteilte Nachrichten, die von der politischen Position einer Person abweichen, nachrangig platziert wurden, sodass sie seltener prominent im Newsfeed erschienen, und Inhalte von den im persönlichen Netzwerk durchaus vorhandenen Menschen mit abweichender politischer Auffassung somit benachteiligt dargestellt wurden.[289] In einer Untersuchung von Google-Suchergebnissen zu Parteien im Vorfeld der Bundestagswahl 2017 konnten hingegen kaum Personalisierungseffekte bei den Ergebnissen festgestellt werden.[290]

Eng mit der Metapher der Filterblase verbunden ist das Konzept der Echokammer, nach dem die algorithmisch geförderte Homogenisierung des personalisierten Angebots dazu führt, dass Überzeugungen und Meinungen zu kontroversen Themen durch ständige Wiederholungen und gegenseitige Bestätigung innerhalb eines Netzwerks Gleichgesinnter verstärkt und verfestigt werden. Dieser Effekt konnte zum Beispiel in einer Studie auf den algorithmisch kuratierten Plattformen Twitter (zum Thema Schwangerschaftsabbrüche) und Facebook (zum Thema Impfen und beim Nachrichtenkonsum) nachgewiesen werden, nicht jedoch beim Nachrichtenkonsum auf der Plattform Reddit, auf der die Reihenfolge von Inhalten

288 Pariser 2011.
289 Bakshy/Messing/Adamic 2015.
290 Krafft/Gamer/Zweig 2018.

nicht algorithmisch personalisiert wird, sondern davon abhängt, wie viele Stimmen ein Beitrag von anderen Teilnehmenden einer Diskussion erhalten hat.[291]

Das Ausmaß und die Wirkung von Filterblasen und Echokammern sind umstritten und dürften sich zwischen verschiedenen Anwendungen und Regionen stark unterscheiden.[292] Es ist also Vorsicht geboten bei zu schnellen Verallgemeinerungen. Ob und in welchem Ausmaß Filterblasen oder Echokammern entstehen oder verstärkt werden, hängt von der verwendeten Technologie, dem Kontext und den Nutzungsweisen ab. Ebenso trägt vermutlich die Wahl der jeweiligen methodischen Ansätze bei der Untersuchung solcher Phänomene dazu bei, dass es zu unterschiedlichen Einschätzungen kommt. Auch der bereits erwähnte mangelnde Zugang zu Plattformdaten für die Forschung erschwert solche wissenschaftlichen Analysen zusätzlich. Allerdings sind die potenziellen Gefahren solcher informationellen Defizite derart essenziell für demokratische Meinungs- und Willensbildungsprozesse, dass ihnen auch bereits bei unklarer Datenlage entschieden entgegengewirkt werden sollte.

7.3.3 Moralische und emotionale Aufladung

Klarer nachweisen lässt sich die zunehmende moralische und emotionale Aufladung des Informationsangebots auf vielen Plattformen. Sie ist zum einen deshalb zu verzeichnen, weil emotionsbehaftete Inhalte grundsätzlich mehr Aufmerksamkeit und Reaktionen auslösen und sich so nach aktuell üblichen Kuratierungskriterien gut verbreiten. Gerade Ärger und Wut verbreiten sich besonders schnell und weit und beziehen sich bei öffentlichen Plattformen häufig auf politische Inhalte.[293]

291 Cinelli et al. 2021.
292 Stark/Magin/Jürgens 2021; Dubois/Blank 2018; Rau/Stier 2019.
293 Fan et al. 2014.

Auf Twitter wurde beispielsweise anhand kontroverser, den US-amerikanischen politischen Diskurs prägender Themen gezeigt, das Tweets umso häufiger geteilt wurden, je mehr moralisch-emotionale Wörter wie zum Beispiel „angreifen", „schlecht" oder „beschuldigen" enthalten waren, und schon ein einziges solches Wort die Verbreitung um 20 Prozent erhöhte.[294] Mit Blick auf die Videoplattform YouTube konnte derweil nachgewiesen werden, dass deren Empfehlungsalgorithmen im Lauf der Zeit zunehmend extreme Inhalte vorschlagen und der Medienkonsum sich auch entsprechend verändert.[295]

Hinzu kommt, dass besonders gefühlsbetonte oder intensive Reaktionen wie beispielsweise ein mehrfaches Hin-und-her-Kommentieren von Algorithmen mitunter auch noch zusätzlich ausdrücklich positiv gewichtet werden. Die 2018 von Facebook eingeführten Neuerungen seiner Gewichtungskriterien, die laut eigenen Angaben dazu gedacht waren, „Gespräche und bedeutungsvolle Interaktionen anzuregen" und „Menschen näher zusammenzubringen"[296], führten beispielsweise dazu, dass fortan Inhalte priorisiert wurden, zu denen es besonders viele emotionale Reaktionen und Kommentare gab und die häufig geteilt wurden. Nach der Einführung der Änderung stieg die Zahl der Reaktionen und Interaktionen auf Facebook schnell erheblich an. Am besten verbreiteten sich allerdings kontroverse Beiträge, auf die mit dem „Wut"-Emoji reagiert wurde oder die hitzige Diskussionen auslösten. Dieser Trend wurde von Personen und Organisationen, die Facebook zur Kommunikation nutzen, schnell erkannt und führte zu einem teilweise bewussten Wechsel hin zur Produktion

294 Brady et al. 2017. Die 15 wirkungsvollsten moralisch-emotional aufgeladenen Worte waren: *attack*, *bad*, *blame*, *care*, *destroy*, *fight*, *hate*, *kill*, *murder*, *peace*, *safe*, *shame*, *terrorism*, *war* und *wrong* (ebd., Appendix Tabelle S3).

295 Ribeiro et al. 2020.

296 https://about.fb.com/news/2018/01/news-feed-fyi-bringing-people-closer-together [12.01.2023].

provokanterer Inhalte, gerade auch im politischen Bereich, da sich nur noch so Reichweite erzielen ließ (vgl. Abschnitt 7.4.1).[297]

7.3.4 Relevanz der beobachteten Effekte

Wie weit verbreitet die hier postulierten Effekte auf die Informationsqualität sind und welche Relevanz dies für politische Prozesse hat, kann derzeit nicht abschließend beantwortet werden. Die Bedeutung von Filterblasen und Echokammern wird, wie bereits angedeutet, kontrovers diskutiert und mit Blick auf Falschinformationen gibt es zahlreiche und teilweise von Plattformen selbst bereitgestellte Werkzeuge und Angebote, um den Wahrheitsgehalt und die Seriosität von Inhalten eigenständig zu prüfen. Dem stehen selbstverständlich zahlreiche positive Auswirkungen der digitalen Transformation für den öffentlichen Meinungsbildungsprozess entgegen, etwa die Möglichkeit, aus einem weitaus größeren und diversen, idealerweise hochwertigen informationellen Angebot zu wählen.

Gleichzeitig erscheint plausibel, dass Falschnachrichten, Filterblasen und Echokammern sowie eine emotional-moralische Zuspitzung vieler Inhalte im regulären Nutzungsalltag auf algorithmisch kuratierten Plattformen negative Auswirkungen auf die Informationsqualität haben können. Dies gilt vor allem in Situationen, in denen Menschen kein spezielles Erkenntnisinteresse verfolgen, sondern vorwiegend die von der Empfehlungssoftware zuvorderst präsentierten Inhalte konsumieren – erst recht, wenn die Kuratierungsmechanismen nicht bekannt sind oder nicht aktiv reflektiert werden.

Hinzu kommt, dass selbst wo ein Bewusstsein für potenziell problematische Effekte vorhanden ist und der Wille besteht, diesen mit dem eigenen Nutzungsverhalten vorzubeugen, die

297 https://www.wsj.com/articles/facebook-algorithm-change-zuckerberg-11631654215 [11.01.2023].

Umsetzung solcher Vorsätze angesichts der psychologischen Wirkmacht vieler Plattformangebote sehr schwierig sein kann. Die Personalisierung des Informationsangebots, seine soziale Einbettung durch Likes, Follower, Kommentare und sonstigen Reaktionen wie auch die beschriebene Priorisierung von Inhalten, die besonders überraschend oder emotional anregend sind, sorgen für die fortlaufende Bereitstellung höchst attraktiver Stimuli für das menschliche Belohnungssystem[298], denen man sich nur schwer entziehen kann[299] und die unsere Aufmerksamkeit überaus erfolgreich[300] und nachhaltig[301] beanspruchen. Die Freiheit, qualitativ hochwertige Informationen zu finden, wird unter diesen Umständen durch die Wirkmacht der zum Einsatz kommenden Algorithmen praktisch vermindert.

7.4 Diskursqualität

Die beschriebenen Änderungen in der Qualität, Darbietung und Verbreitung algorithmisch vermittelter Informationen verändern auch die Qualität vieler online stattfindender Diskurse in ethisch wie politisch relevanter Hinsicht. Auch hier sind zunächst wieder positive Entwicklungen und Potenziale zu benennen, die sich insbesondere aus den auf Plattformen und in sozialen Medien wesentlich erhöhten Möglichkeiten zu Teilhabe und direkter Vernetzung ergeben. Angesichts der weltweit zunehmenden Verfügbarkeit von Onlinezugängen und internetfähigen Endgeräten ist die Schwelle zur Beteiligung an öffentlichen Diskursen für viele Menschen deutlich gesunken. Die Kuratierungskriterien auf Plattformen und in sozialen Medien geben im Gegensatz zu traditionellen Medien auch weniger etablierten Akteuren und Institutionen die

298 Burhan/Moradzadeh 2020.
299 He/Turel/Bechara 2017.
300 Ward et al. 2017.
301 Mickes et al. 2013.

Chance, mit gut formulierten Beiträgen sehr viele Personen direkt zu erreichen und – je nach Plattform – auch von einflussreichen Individuen und Organisationen direkt beachtet zu werden. Im Journalismus beispielsweise wurden aufgrund der erweiterten Möglichkeiten zur direkten Kontaktaufnahme mitunter wesentliche Verbesserungen in der Diskursqualität wahrgenommen.[302]

Manche Diskurse erhalten durch die auf Plattformen und sozialen Medien geschaffenen soziotechnischen Infrastrukturen überhaupt erst einen Raum, weil es über diese Onlineangebote möglich wird, sich mit Menschen mit ähnlichen Interessen oder Problemen ortsunabhängig auszutauschen und zu vernetzen, auch zu sehr speziellen Themen, zu denen sich lokal und offline vielleicht kaum Gelegenheiten zum Diskurs ergäben. Durch solche Zusammenschlüsse können zudem Anliegen, die sonst wohl wenig Chancen auf Wahrnehmung hätten, auch in der breiteren Öffentlichkeit Beachtung finden. Auch Potenziale für die Auseinandersetzung mit Argumenten und anderen Positionen werden durch die neuen Diskursmöglichkeiten zunächst einmal erweitert. Wer dies möchte, kann mit entsprechenden Suchanfragen und Interaktionen auf Plattformen und in sozialen Medien die eigenen Horizonte erweitern, Potenziale für Perspektivenwechsel entdecken und mehr Verständnis und Empathie für Andersdenkende entwickeln.

Inwieweit solche Möglichkeiten sich in der Praxis entfalten können, wird allerdings kontrovers beurteilt, denn gegenüber den genannten Chancen werden auch mit Blick auf die Diskursqualität negative Entwicklungen diskutiert. Dabei geht es vor allem um drei Themen. Das erste betrifft die Frage, ob und inwieweit Plattformen und soziale Medien zu einer in den letzten Jahren beobachteten zunehmenden politischen Polarisierung öffentlicher Diskurse beitragen. Das zweite Thema berührt Praktiken der politischen Werbung und Manipulation

302 https://www.theguardian.com/technology/2022/nov/30/twitter-journalism-elon-musk [12.01.2023].

und wie diese sich aufgrund der neuen, algorithmisch vermittelten Kommunikationsmöglichkeiten verändern. Ein dritter Diskussionsschwerpunkt gilt Kontroversen um Diskursverrohungen mit Hassrede und ähnlichen Integrationsverletzungen auf der einen und überbordenden Eingriffen in die Äußerungs- und Meinungsfreiheit auf der anderen Seite.

Alle drei Themen beschreiben komplexe Phänomene, die es zweifelsohne nicht erst seit Beginn der digitalen Transformation des öffentlichen Diskurses gibt und die sowohl multifaktoriell bedingt als auch in ihrem Facettenreichtum oft schwer einheitlich fassbar sind. Indizien, dass sich aktuelle Trends in der technisch mediierten öffentlichen Kommunikation über Plattformen und soziale Medien in diesen Bereichen auf die Diskursqualität auswirken, sind vor dem Hintergrund dieser Komplexität zu betrachten und sollten daher nicht zu vorschnellen Schlüssen über ursächliche Zusammenhänge verleiten.

7.4.1 Politische Polarisierung

Es gibt zahlreiche Hinweise, dass der in Abschnitt 7.3.1 beschriebene Umstand, dass emotional und moralisch aufgeladene Inhalte sich auf Plattformen besonders schnell und weit verbreiten, zu Tonfallverschiebungen geführt hat, auch und insbesondere auf vielen Kanälen, die aktiv zur Gestaltung des politischen Diskurses beitragen. Im Zuge der Enthüllungen der „Facebook Files" durch das Wall Street Journal wurde ab September 2021 eine Reihe von Belegen veröffentlicht[303], dass Kommunikationsteams im Bereich Politik sich nach den 2018 eingeführten Änderungen im Facebook-Algorithmus auch gezielt zunehmend provokanteren und negativen Inhalten und Kommunikationsstilen zuwandten und dass Facebook sich dessen bewusst war (siehe Infokasten 8).

303 https://www.wsj.com/articles/the-facebook-files-11631713039 [11.01.2023].

Infokasten 8: „The Facebook Files" – Verminderung der Kommunikationsfreiheit durch algorithmische Formatierung

Ab September 2021 veröffentlichte das Wall Street Journal eine Artikelserie, die sogenannten „Facebook Files", in der auf Grundlage von Informationen der Whistleblowerin Frances Haugen zahlreiche Details darüber enthüllt wurden, in welchem Ausmaß Facebook gesellschaftlich bedenkliche Auswirkungen seiner Produkte bekannt waren. Dabei ging es auch darum, wie die im Jahr 2018 vorgenommenen Änderungen der Facebook-Algorithmen dazu beigetragen haben, den politischen Diskurs zu polarisieren.[304]

Nach der Einführung neuer Gewichtungskriterien, die ursprünglich zur Förderung „bedeutungsvoller Interaktionen" konzipiert worden waren, stellte sich bald heraus, dass fortan vor allem solche Inhalte erfolgreich waren, auf die Menschen besonders aufgebracht reagierten, etwa weil sie besonders provokant oder schockierend wirkten oder Themen behandelten, zu denen es ausgeprägte emotionale Kontroversen gab (vgl. Abschnitt 7.3.3). Diese Verschiebung fiel bald auch denjenigen auf, die über Facebook einen Großteil ihrer öffentlichen Kommunikation abwickelten und daher darauf angewiesen waren, über die Plattform Reichweite zu erzielen.

Gerade im politischen Bereich gab es laut dem internen Facebook-Bericht „Political Party Response to '18 Algorithm Change" entsprechende Beobachtungen und daran angepasste Änderungen in der Kommunikationsstrategie hin zu „provokativen Inhalten von niedriger Qualität" (im englischen Original: „provocative, low-quality content"). Die Reichweite positiver und sachlicher Beiträge hatte sich nach Angaben aus den Kommunikationsteams politischer Parteien aus verschiedenen Ländern deutlich reduziert, während hetzerische Inhalte und direkte Angriffe auf die politische Konkurrenz erfolgreich liefen. Das Social-Media-Team einer polnischen Partei beispielsweise schätzte, dass es den Tenor seiner Inhalte von einer zuvor ausgewogenen Balance zwischen positiven und negativen Beiträgen infolge der neuen algorithmischen Auswahlkriterien zu 80 Prozent in Richtung negative Beiträge verändert hatte, um weiterhin Reichweite zu erzielen. Ähnliche Angaben machten dem Bericht zufolge unter anderem Parteien aus Spanien, Taiwan und Indien, verbunden mit Beschwerden aus den Kommunikationsteams, die diese Entwicklung als negativ und demokratiegefährdend einschätzten – sich aber gleichwohl außerstande sahen, sich ihr zu entziehen.

Erwähnenswert ist weiterhin, dass die Geschäftsleitung einschließlich Mark Zuckerberg nach den Belegen der Whistleblower auch nach Offenlegung der bedenklichen Zusammenhänge Maßnahmen zu deren Entschärfung mit Verweis auf die Priorisierung ökonomischer Interessen ablehnte.

Hieran knüpfen sich zahlreiche Fragen zur Verantwortung von Internetfirmen für das Gemeinwohl, die im weiteren Text noch näher beleuchtet werden.

304 https://www.wsj.com/articles/facebook-algorithm-change-zuckerberg-11631654215 [11.01.2023]. Alle weiteren Angaben in diesem Infokasten stammen aus diesem Artikel.

Durch die algorithmische Konfigurierung und die sich daraus ergebenden Wechselwirkungen zwischen Menschen und algorithmischen Systemen ergaben sich hier also für die am Diskurs teilnehmenden Personen und Organisationen verminderte Handlungsspielräume mit negativen Auswirkungen auf die Diskursqualität. Aus den Facebook-Dokumenten geht zudem hervor, dass die Auswirkungen von Änderungen in den Algorithmen auch für die intern bei Facebook an den Programmier- und Entscheidungsprozessen Beteiligten mitunter schwer abzuschätzen waren. Dies unterstreicht die Komplexität der online entstandenen soziotechnischen Systeme, die zu überraschenden Rückwirkungen auf menschliches Handeln führen kann.

Die Beobachtung, dass auf Polarisierung ausgelegte Inhalte erfolgreicher sind, sendet deutliche Signale an alle, die politisch kommunizieren. Erhöhte Chancen auf virale Verbreitung durch die Verwendung einer moralisierenden Sprache, die Empörung über die politische Gegenseite auslöst[305], fallen auch denjenigen auf, die für Kommunikationsstrategien verantwortlich sind und den Erfolg ihrer Kampagnen analysieren. Die Möglichkeit, mithilfe der von den Plattformen bereitgestellten Analysetools sehr schnell Rückmeldung zur Effektivität der eigenen Kommunikation zu erhalten, fördert die ständige Optimierung dieser Effektivität nach den Kriterien der Plattform und treibt selbst nach Einschätzung von Personen, die aktiv an solchen Optimierungsversuchen mitwirken, eine ungünstige Polarisierung des politischen Diskurses voran.[306]

Eine postulierte Spaltung der Öffentlichkeit findet angesichts dieser Beobachtungen womöglich weniger deswegen statt, weil mögliche Filterblaseneffekte Inhalte und Argumente von außerhalb des personalisierten Präferenzenprofils ausschließen und zu fragmentierten Diskursen führen; vielmehr

305 Für Twitter vgl. etwa Brady et al. 2017.
306 https://www.humanetech.com/podcast/53-how-political-language-is-engineered [11.01.2023].

führt nach dieser Deutung die Priorisierung emotional und moralisch aufgeladener Kommunikation gerade zu einer ständigen Konfrontation mit Inhalten und Argumenten von als gegnerisch wahrgenommenen Gruppen, die auf Grundlage des eigenen personalisierten Präferenzenprofils als bedrohlich, verwerflich oder anderweitig abzulehnend präsentiert werden.[307] Eine verschärfte Polarisierung und Radikalisierung von Positionen und Befindlichkeiten geht demnach vor allem auf diese andauernde Reibung zurück, mit dem Ergebnis, dass die Bereitschaft zum offenen, rationalen und wohlwollenden Diskurs stark eingeschränkt wird und die Spielräume für demokratisches Zusammenwirken im Sinne eines gemeinsamen öffentlichen Vernunftgebrauchs sich damit vermindern.

Solche unterschiedlichen Erklärungsansätze können auch auf der empirischen Ebene zu unterschiedlichen und mitunter widersprüchlichen Resultaten führen. Um möglichen Mechanismen genauer auf den Grund zu gehen, erscheint daher auch in diesem Kontext mehr multimethodische Forschung empfehlenswert, für die auch in diesem Fall wiederum ein besserer Zugang zu Plattformen und ihren Daten notwendig wäre.

Vorerst bleibt aufgrund der noch unzureichenden Studienlage umstritten, welche Rolle algorithmisch mediierte Kommunikationsprozesse für eine beobachtete oder vermutete gesellschaftliche Diskursverschärfung spielen. Angesichts vielfältiger weiterer Einfluss- und Änderungsfaktoren, die auf politisch relevante Situationen, Prozesse und Befindlichkeiten wirken können, gibt es sicherlich zahlreiche weitere Ursachen hierfür, die nicht primär an der besonderen Qualität technisch vermittelter Diskurse liegen. Sie können jedoch im Rahmen solcher Diskurse aufgegriffen werden und Verstärkereffekte erfahren.

307 Törnberg 2022.

7.4.2 Politische Werbung und Manipulation

Angesichts der zunehmenden Bedeutung von sozialen Medien und Plattformen bei politischen Auseinandersetzungen stehen neben den beschriebenen Polarisierungstendenzen auf offiziellen Kanälen auch weitere, mitunter im Verborgenen laufende Aktivitäten zur Beeinflussung politischer Prozesse in der Diskussion. Nun sind Versuche, die öffentliche Meinung durch wahrhaftige und unwahrhaftige, wahre und falsche Behauptungen und Stellungnahmen zu verändern, ein uraltes Phänomen der Politik, das schon für die antiken Vorformen der modernen Demokratie in Griechenland und Rom eine wichtige Rolle spielte. Auch damals war nicht immer klar, wer mit welchen Absichten welche Behauptungen in die Welt gesetzt hatte und warum sich diese und nicht jene zu einer umfassenden Verleumdungskampagne auswuchs. Durch die Digitalisierung der Kommunikation verschärft sich die Problematik allerdings insofern, als die Zahl kommunikativer Akte nicht begrenzt ist und ihre Kosten minimal sind. So erweitern sich einerseits die Handlungs- und Gestaltungsmöglichkeiten für diejenigen, die an politischen Kommunikationskampagnen mitwirken. Kombiniert mit der Funktionsweise sozialer Medien und Plattformen, die sowohl eine besonders genaue Personalisierung von Inhalten und Überprüfung ihrer Effektivität ermöglicht als auch die Verbreitung polarisierender, persuasiver, emotionsbetonter und moralisierender Botschaften fördert, ergibt sich zudem viel Potenzial für besonders wirkmächtige Kommunikationskampagnen, die eingebettet in den digitalen Alltag ablaufen, ohne dass die von solchen Kampagnen erreichten Personen sich dessen gewahr werden. Dies vermindert jedoch andererseits die Freiheit der von diesen Kampagnen adressierten Personen, politische Werbung zu erkennen und sich bewusst damit auseinanderzusetzen.

Die in diesem Zusammenhang relevanten Entwicklungen gehen auf die gleichen Mechanismen zurück, die bereits in Abschnitt 7.2 als grundlegend für die Funktionsweise und das

Geschäftsmodell sozialer Medien beschrieben wurden. Auf Basis der reichhaltigen datenbasierten Profile, die sich aus den digitalen Spuren, die man bei der Nutzung von Plattformen hinterlässt, erstellen lassen, kann man nicht nur konsumrelevante Interessen extrahieren, um das Schalten personalisierter kommerzieller Werbung zu ermöglichen, sondern auch psychologische Merkmale und politische Neigungen sehr präzise ableiten.[308] Hier besteht die Gefahr, dass solche Informationen manipulativ eingesetzt werden, um beispielweise zielgenaue politische Werbung zu schalten (Targeted Advertisement) oder um die Wählerschaft, deren Interessen der politischen Konkurrenz zuneigen, strategisch zu desinformieren oder von der Wahl abzuhalten. Unternehmen wie die bis 2018 operierende Datenanalyse- und Beraterfirma Cambridge Analytica verfolgen ausdrücklich das Ziel, aus dem Nutzungsverhalten und sonstigen Datenspuren politische Persönlichkeits- und Interessenprofile zu erstellen und mit deren Hilfe für bestimmte Gruppen oder Individuen maßgeschneiderte Botschaften zu verbreiten (Microtargeting), um politische Meinungsbildungsprozesse und Wahlen zu beeinflussen (siehe Infokasten 9).

Infokasten 9: Cambridge Analytica

Das Unternehmen Cambridge Analytica wurde 2013 gegründet, um politische Kampagnen mit einer Mischung aus datenbasierter Verhaltensforschung und strategischer Kommunikation zu unterstützten. Seine Kundschaft kam vornehmlich aus dem konservativen Spektrum, darunter die damaligen US-Präsidentschaftskandidaten Donald Trump und Ted Cruz. Die Firma beanspruchte, mit ihren Algorithmen besonders präzise psychologische Einschätzungen von Einzelpersonen aus deren Nutzungsdaten ableiten zu können und diese für einen psychografisch personalisierten Zuschnitt politischer Botschaften nutzbar zu machen – ein Ansatz, dessen Wirksamkeit nie eindeutig nachgewiesen und von Cambridge Analytica vermutlich überspitzt dargestellt wurde, um das Interesse am Unternehmen zu schüren.

Die Firma verwendete in ihren Kampagnen aber auch Methoden, die ohne die Berücksichtigung aufwendiger psychografischer Profile auskommen, sondern stattdessen schlicht die Wahlmotivation bestimmter ethnischer

308 Kosinski/Stillwell/Graepel 2013; Matz/Appel/Kosinski 2020.

oder demografischer Gruppen zu senken versuchten. Im US-Präsidentschaftswahlkampf 2016 beispielsweise nutzte Cambridge Analytica Informationen aus Facebook-Profilen, um afroamerikanischen Menschen, von denen auf Grundlage vorheriger Wahlergebnisse zu erwarten war, dass sie ihre Stimme mit hoher Wahrscheinlichkeit den Demokraten geben, gezielt negative Inhalte über Trumps Konkurrentin Hillary Clinton anzuzeigen, beispielsweise ein Videoclip aus dem Jahr 1996, in dem sie kriminelle afroamerikanische Jugendliche angeblich als „Superraubtiere" bezeichnete.[309] Besonders der Umstand, dass diese Kampagnen oft ganz oder teilweise im Verborgenen abliefen, ohne dass Auftraggebende die Beauftragung offenlegten oder Betroffene wussten, dass sie Ziele personalisierter politischer Werbung waren, wird als demokratiegefährdend gewertet.[310]

2018 gab der Whistleblower Chris Wylie in Interviews bekannt, dass große Teile der Datenbasis von Cambridge Analytica missbräuchlich ohne Einwilligung der Nutzenden beschafft worden waren. Hierfür war eine Fragebogen-App für Facebook-Nutzende entwickelt worden, die nicht nur die Daten von Teilnehmenden sammelte, sondern unbekannterweise auch von den Profilen aus ihrem persönlichen Netzwerk.[311] Dies löste eine breite öffentliche Diskussion über Sicherheit und Zugang zu den Daten in sozialen Netzwerken sowie die Legitimität personalisierter politischer Werbung aus. 2020 folgten weitere Enthüllungen von der ehemaligen Cambridge-Analytica-Mitarbeiterin Brittany Kaiser, die über Twitter zahlreiche Dokumente veröffentlichte, die eine Einflussnahme des Unternehmens auf Wahlen und Politik in 68 Ländern belegten.[312]

Wie erfolgreich personalisierte politische Werbung mit Microtargeting-Ansätzen tatsächlich sein kann, ist zwar umstritten bzw. noch nicht hinreichend erforscht.[313] Es gibt jedoch deutliche Hinweise darauf, dass schon die akzentuierte Konfrontation mit bestimmten Inhalten politische Meinungen beeinflussen kann. In Experimenten konnte beispielsweise gezeigt werden, dass die Reihenfolge, in der Suchmaschinen Ergebnisse präsentieren, wenn politisch noch unentschiedene Personen online Informationen über Kandidatinnen und Kandidaten suchen, entscheidend auf die Meinungsbildung

309 https://www.channel4.com/news/revealed-trump-campaign-strategy-to-deter-millions-of-black-americans-from-voting-in-2016 [11.01.2023].
310 Kaiser 2020.
311 https://www.theguardian.com/news/2018/mar/17/data-war-whistleblower-christopher-wylie-faceook-nix-bannon-trump [11.01.2023].
312 https://www.theguardian.com/uk-news/2020/jan/04/cambridge-analytica-data-leak-global-election-manipulation [16.01.2023].
313 Bodó/Helberger/de Vreese 2017.

auswirkt, da weiter oben platzierte Ergebnisse als bedeutsamer eingestuft werden.[314] Allein das Wissen darum, dass versucht wird, auf Grundlage sehr persönlicher psychologischer Merkmale politische Präferenzen zu manipulieren, kann zudem plausibel negative Effekte auf den politischen Diskurs und das Vertrauen in politische Meinungsbildungsprozesse entfalten.

Vertrauensschädigend kann sich weiterhin der Umstand auswirken, dass zur strategischen Beeinflussung des öffentlichen politischen Diskurses auch vielfach unechte Profile (Fake-Accounts) eingesetzt werden, die teilweise automatisiert betrieben werden (Bots). In einer investigativen Recherche trug die New York Times hierzu vielfältige Daten zusammen, nach denen beispielsweise bis zu 27 Prozent der Beiträge auf Facebook und Twitter im Zusammenhang mit den Wahlen in Mexiko im Jahr 2018 von Bots und Fake-Accounts stammten[315] und Chinas staatliche Nachrichtenagentur Xinhua hunderttausende falsche Profile und Retweets bezahlte, die Inhalte gezielt an westliche Twitter-Konten schicken[316]. In einer Anhörung vor dem US-Senat zum Einfluss Russlands auf die Präsidentschaftswahl im Jahr 2016 legten die Vertreter von Facebook, Twitter und Google offen, dass etliche Personen in den USA über künstliche Profile mit Inhalten der russischen „Agentur für Internetforschung" Glavset konfrontiert worden waren, darunter ca. 150 Millionen Personen allein über Inhalte auf Facebook und Instagram.[317]

Auch die Wirkmacht von Fake-Accounts und Bots auf politische Prozesse ist umstritten und lässt sich oft schwer präzise fassen. Schon die Identifikation unechter Profile stellt eine sich ständig fortentwickelnde Herausforderung dar.[318] Es gibt jedoch zahlreiche Hinweise, dass Kommunikationskampagnen

314 Epstein/Robertson 2015.
315 https://www.nytimes.com/2018/05/01/world/americas/mexico-election-fake-news.html [12.01.2023].
316 https://www.nytimes.com/interactive/2018/01/27/technology/social-media-bots.html [12.01.2023].
317 Stretch, in: United States Senate 2018, 58.
318 Ferrara et al. 2016.

über Fake-Accounts zumindest das Potenzial für eine hohe Effektivität haben. So demonstrierte eine experimentelle Studie, dass wenige strategisch gut platzierte Bots, die extreme Botschaften verbreiten, nach aktuellen Kuratierungskriterien überdurchschnittlich erfolgreich auf Entscheidungen Einfluss nehmen können.[319] In einer Untersuchung von über 240 Millionen Tweets zur US-Präsidentschaftswahl im Jahr 2020 wurde deutlich, dass wenige tausend Bots genauso viele Spitzen im Kommunikationsgeschehen zu aktuellen politischen Themen verursachten wie die Konten echter Menschen, und das vorwiegend, indem sie lediglich über Retweets als Verstärker bestimmter Botschaften wirkten und ihnen somit mehr Überzeugungskraft verliehen.[320] Mit solchen Verstärkereffekten können Diskurse problematisch verzerrt werden, sodass sie nicht mehr das Interesse oder die Ansichten der an ihnen beteiligten Menschen spiegeln. Verunsicherungen und der Verlust des Vertrauens in den Diskursprozess selbst können die Folge sein, wenn die Legitimität der Diskurse und die Identität der Teilnehmenden zunehmend bezweifelt werden.

Dass solche Zweifel berechtigt sein können, zeigen Analysen, die nachweisen, dass mit unechten Profilen im Rahmen der Funktionsweise eines sozialen Netzwerks auch langfristige Strategien zur subtilen Beeinflussung verfolgt werden. Hier werden beispielsweise zunächst über längere Zeiträume Reichweite und Netzwerke mit unpolitischen Inhalten wie humorvollen Memes (Text- und Bildmontagen) zu identitätsstiftenden Themen wie Familie, Mode oder lokale Kultur aufgebaut. Erst später oder nur punktuell werden Botschaften solcher Konten um politische Inhalte angereichert, in oft subtiler Weise und stark aufbauend auf zuvor etablierten Gruppenidentitäten.[321]

319 Stewart et al. 2019.
320 Ferrara et al. 2020.
321 DiResta et al. 2019; Bradshaw/DiResta/Miller 2022; Bradshaw/Henle 2021.

7.4.3 Spannungsfeld Diskursverrohung und Äußerungsfreiheit

Im Zusammenhang mit den beschriebenen Beobachtungen zur Verschärfung von Tonlagen auf Plattformen und in sozialen Medien wird häufig die Annahme vorgebracht, dass dieser Trend auch zu einer Verrohung des politischen Diskurses beitragen kann. Besorgniserregend ist hier insbesondere die Zunahme stark negativ und aggressiv geprägter Kommunikationsstile bis hin zu Hassrede, Drohungen und Aufforderung zu Gewalt. Regelmäßige Untersuchungen der Landesanstalt für Medien NRW zur Hassrede im Internet zeigen, dass die Wahrnehmung von Hasskommentaren seit 2016 zugenommen hat und seit 2018 auf unverändert hohem Niveau verharrt.[322] Ansätze zur Moderation von Inhalten hinken dieser Entwicklung häufig hinterher. Laut den vom Wall Street Journal in den „Facebook Files" dargestellten Enthüllungen schätzen Facebook-Angestellte beispielsweise den Erfolg ihrer Bemühungen, Hassbotschaften mit algorithmischen Methoden zu identifizieren und einzudämmen, als überaus unzureichend ein, mit Erfolgsquoten im niedrigen einstelligen Bereich.[323]

Die bloße Existenz solcher Inhalte ist bereits für sich genommen problematisch, da Gehässigkeit, Verleumdungen und Einschüchterungen auf sozialen Medien so viel Unbehagen und Angst schüren können, dass dies Personen davon abhält, ihre Meinung zu äußern, im Internet präsent zu sein oder sich am öffentlichen Diskurs zu beteiligen. Man spricht auch hier von Chilling-Effekten, die diskriminierend auf Betroffene wirken und deren Freiheit und Handlungsmöglichkeiten in der Onlinekommunikation erheblich vermindern können (vgl. Abschnitt 6.4).

322 Landesanstalt für Medien NRW 2020; 2022.
323 https://www.wsj.com/articles/facebook-ai-enforce-rules-engineers-doubtful-artificial-intelligence-11634338184 [12.01.2023].

Hinzu kommt die Gefahr, dass online verbreitete Hetze in bedrohliche Handlungen in der realen Welt umschlägt. Auch wenn kausale Zusammenhänge im Einzelfall schwer nachweisbar sein mögen, gibt es genügend Korrelationen zwischen solchen Vorfällen, sodass diese Sorge gerechtfertigt erscheint und jedenfalls plausibel zu den genannten Chilling-Effekten beiträgt. So wurden die 2017 an den Rohingya in Myanmar verübten Massaker durch zahlreiche Hassbotschaften und Gewaltaufrufe auf Facebook befeuert, das zu diesem Zeitpunkt aufgrund seiner kostenlosen Nutzbarkeit über die Mobilfunknetze monopolartig den Zugang der Bevölkerung zum Internet darstellte. Die Bemühungen des Konzerns, diese problematischen Inhalte zu moderieren, wurden auch hier weitgehend als unzureichend eingestuft.[324] Ähnliche Vorwürfe über die Rolle Facebooks bei ethnisch motivierter Gewalt werden beispielsweise im Zusammenhang mit dem Bürgerkrieg in Äthiopien[325] und religiös motivierter Gewalt in Indien[326] diskutiert.

In einer westlichen Demokratie wurde das demokratiegefährdende Potenzial von online verbreiteten Entrüstungsstürmen am deutlichsten augenfällig, als in den USA im Januar 2021 zahlreiche Personen, die die Niederlage des abgewählten Präsidenten Donald Trump nicht anerkennen wollten, nach Befeuerung in den sozialen Medien – unter anderem durch Trump selbst – das Kapitol stürmten, um den Senat und das Repräsentantenhaus an der förmlichen Bestätigung des Wahlsieges von Joe Biden bei der Präsidentschaftswahl 2020 zu hindern. Eine Reaktion auf dieses Ereignis, bei dem fünf Menschen starben und zahlreiche weitere verletzt wurden, war die Deaktivierung von Trumps Accounts auf Facebook und Twitter durch die entsprechenden Plattformbetreiber. Die

324 https://www.reuters.com/investigates/special-report/myanmar-facebook-hate [12.01.2023].
325 https://www.theguardian.com/technology/2022/feb/20/facebook-lets-vigilantes-in-ethiopia-incite-ethnic-killing [12.01.2023].
326 https://www.wsj.com/articles/facebook-services-are-used-to-spread-religious-hatred-in-india-internal-documents-show-11635016354 [12.01.2023].

internen Entscheidungsprozesse bei Twitter zur Moderation von Inhalten und Konten im Umfeld dieser Ereignisse wurden nach der Übernahme des Konzerns durch Elon Musk im Rahmen der „Twitter Files" ab Dezember 2022 veröffentlicht und teils kritisch beurteilt.[327] Solche Reaktionen werfen ihrerseits demokratietheoretische Fragen auf. Müssen und sollen wir es hinnehmen, dass große private Anbieter, die teilweise eine monopolistische Stellung auf den Medienmärkten haben, entscheiden, wer in den sozialen Medien zu Wort kommen kann und wer nicht? Wollen wir, dass Community-Standards, die von privaten Medien intern festgelegt werden, die Formen der Kommunikation auch jenseits rechtlicher Vorgaben bestimmen?

Der Umgang mit solchen Fragen gestaltet sich schwierig, da hier zwei normativ relevante Gesichtspunkte einander gegenüberstehen. Auf der einen Seite können übermäßige Löschungen und Sperrungen (Overblocking) einen Eingriff in die Meinungs- und Pressefreiheit darstellen. Overblocking kann selbst zu Chilling-Effekten beitragen, nämlich dann, wenn Menschen bestimmte Inhalte gar nicht erst veröffentlichen, weil sie befürchten, dass diese gleich wieder gelöscht werden und ihrem Konto Einschränkungen drohen. Bei solchen Eingriffen wird auch die Frage relevant, *wer* über Löschungen bzw. deren Kriterien entscheidet. Bereits die Löschpraktiken sozialer Medien sind solche Einschnitte. Meinungs- und Pressefreiheit sind verfassungsrechtlich in erster Linie Abwehrrechte gegenüber dem Staat, denen das Hausrecht privater Plattformen nicht unterworfen ist. Das bedeutet, dass Plattformen durchaus mehr löschen dürfen und können als nur illegale Inhalte. Nach Auffassung des Bundesgerichtshofs sind die Plattformen aber auch bei der Anwendung der eigenen Regeln, die sie über die Vertragsbeziehungen mit ihrer Kundschaft vermitteln, an die Grundrechte der Personen gebunden, die ihr Angebot nutzen,

327 https://www.racket.news/p/capsule-summaries-of-all-twitter [31.01.2023].

etwa deren Freiheit zur Meinungsäußerung.[328] Mit Inkrafttreten des Netzwerkdurchsetzungsgesetzes sind sie allerdings seit Januar 2018 verpflichtet, solche illegalen Inhalte *nach* Meldung durch Dritte zu löschen.[329] Auf der anderen Seite stellt sich jedoch die Frage, wie man, wenn man solche Eingriffe in die Meinungsfreiheit durch Löschungen und Sperrungen ablehnt, verhindern will, dass Hassbotschaften, Aufrufe zu Gewalt, Verleumdungen und sonstige Einschüchterungen die oben beschriebenen Chilling-Effekte aus Angst vor Übergriffen auslösen, Gewalt befördern und weitere demokratiegefährdende Wirkung entfalten. Die gebotene Pluralismussicherung widerstreitet jedoch ebenso vorschnellen algorithmenbasierten Exklusionsmechanismen unter einseitiger Berufung auf angeblich vertrauenswürdige Quellen. An sich ungewollten Chilling-Effekten, die sich aus unter anderem durch gesetzliche Regelungen (vor allem das Netzwerkdurchsetzungsgesetz) bedingten, zu weit reichenden Blockierpraktiken ergeben, ist konsequent entgegenzuwirken.

Ein möglicher Lösungsansatz sieht vor, soziale Medien und Plattformen, soweit diese rundfunkartige Funktionen übernehmen (z. B. der Newsfeed von Facebook), in diesen Funktionen anderen Rundfunkanbietern gleichzustellen, sie also an die einschlägigen gesetzlichen Bestimmungen für den Rundfunk zu binden. Die Kuratierung von Inhalten würde dann auf Grundlage einer rechtlichen Regelung erfolgen, die sicherstellt, dass das Recht auf Meinungsfreiheit gewahrt wird, ohne die demokratische Zivilkultur zu gefährden. Zudem könnte die Etablierung einer öffentlich verantworteten Infrastruktur digitaler Kommunikation in der Europäischen Union als Alternative zu privaten und ausschließlich kommerzgetriebenen

328 Urteile vom 29. Juli 2021 – III ZR 179/20 und III ZR 192/20 (NJW 2021, 3179).

329 Die Europäische Union hat 2022 den Digital Services Act erlassen; Anbieter müssen die neuen Reglungen bis zum 17. Februar 2024 schrittweise umgesetzt haben. Derzeit prüft die Bundesregierung, inwieweit das Netzwerkdurchsetzungsgesetz mit Geltung des Digital Services Act unanwendbar wird, und plant, Materien, die weiterhin nationalstaatlich geregelt werden können, in einem neuen Digitale-Dienste-Gesetz zu bündeln.

Angeboten einen Beitrag zur demokratischen Zivilkultur in Zeiten der digitalen Transformation leisten.

7.4.4 Erweiternde und vermindernde Rückwirkungen auf den öffentlichen Vernunftgebrauch

Die genaue Wirkmacht der hier vorgestellten Veränderungen in der Diskursqualität, die sich aus den spezifischen soziotechnischen Merkmalen und Möglichkeiten öffentlicher Kommunikation auf Plattformen und in sozialen Medien ergeben, ist aktuell noch nicht vollständig abschätzbar. Die vielfältig erweiterten Möglichkeiten zur Teilhabe und Vernetzung sowie grenzüberschreitend wirksamen Kommunikation eröffnen zahlreiche Chancen zur Verbesserung der Diskursqualität und zur Stärkung demokratischer Prozesse.

Die Entfaltung dieser Potenziale scheint in der Praxis jedoch oft schwer planbar zu sein oder wird durch negative Aspekte der Onlinekommunikation behindert. Aufgrund der komplexen Vernetzung von Kommunikation im Internet sind viele politische Diskursprozesse von Einzelnen, aber auch von großen Gruppen nur begrenzt beeinflussbar. Manipulationen und Gegenmanipulationen, die strategischen Formen der Kommunikation[330], schaukeln sich zu chaotischen Prozessen hoch. Diese lassen der politischen Deliberation, der Abwägung von Gründen pro und kontra, womöglich nur noch wenig Spielraum und vermindern so die Möglichkeiten des öffentlichen Vernunftgebrauchs. Extremere und emotionalere Aussagen haben unabhängig von ihrem Wahrheitsgehalt gute

330 Vgl. dazu Jürgen Habermas (1981), der in seiner „Theorie des kommunikativen Handelns" drei Typen rationalen Handelns vorstellt: instrumentelles Handeln, strategisches Handeln und kommunikatives Handeln, sowie das Gibbard-Satterthwaite-Theorem (Gibbard 1973; Satterthwaite 1975; Nida-Rümelin 2020b, 140 ff., 149 ff.).

Chancen, sich gegenüber abgewogenen Argumenten durchzusetzen oder diese sogar zu marginalisieren.

Speziell mit Blick auf Fake-Accounts und Bots kommt hinzu, dass die hier eingesetzte kalkulierte Verzerrung und bewusste Täuschung selbst Rückwirkungen auf die Diskursqualität und Prozesse politischer Willensbildung haben können, indem sie zu Vertrauensverlusten führen. Wenn gefälschten Identitäten in die menschliche Interaktion eintreten, agieren sie auch als Substitute für menschliche Subjekte. Die Zurechenbarkeit einer Botschaft zu einem personalen Gegenüber ist in der Folge nicht mehr gegeben und insbesondere die Delegation kommunikativer Akte an Software kann zu Verunsicherungen führen, wenn man nicht unterscheiden kann, ob man mit Menschen oder Bots kommuniziert. Sowohl das Vertrauen in den Diskurs und die mögliche Einigung auf Kompromisse als auch das Vertrauen in demokratische Prozesse insgesamt, die eigene Urteilskraft sowie die individuellen Wirkmöglichkeiten können auf diese Weise vermindert werden.

7.5 Fazit und Empfehlungen

Die hier aufgezeigten Phänomene und Entwicklungen, die sich in den jungen soziotechnischen Infrastrukturen digitaler Netzwerke vollziehen, haben erhebliche Auswirkungen auf Prozesse der öffentlichen Kommunikation sowie der politischen Meinungs- und Willensbildung, auch und vielleicht insbesondere in demokratischen Gesellschaften. Demokratie als kollektive Selbstbestimmung der Freien und Gleichen setzt individuelle Autonomie voraus. Sie gründet auf dem in Kapitel 3 entwickelten Gedanken, dass Freiheit und Vernunft eng verbunden sind. Praktische Vernunft, die Fähigkeit, das eigene Leben, seine Einstellungen und Handlungen von Gründen leiten zu lassen, ist eine Grundlage von Autorschaft und Voraussetzung für gelingende demokratische Prozesse, die auf rationalen Austausch mit anderen und die freie Auseinandersetzung

unterschiedlicher Auffassungen setzen, um auf dieser Grundlage gute Entscheidungen gemeinsam und wohlbegründet treffen zu können.

Diese Diskurs- und Entscheidungskultur hat im modernen öffentlichen Raum der Plattformen und sozialen Medien vielfältige Entwicklungen erfahren. Durch den enorm verbesserten Zugang zu vielfältigen und auch qualitativ hochwertigen Informationen wie auch durch die stark erhöhten Vernetzungsmöglichkeiten haben sich viele erweiternde Potenziale für einen umfassenderen, sachlich gestützten und dynamischen Austausch ergeben. Menschen, die vorher nur wenig oder keinen Zugriff auf Informationen und Expertise hatten und kaum über Möglichkeiten verfügten, sich über ihr lokales Umfeld hinaus Gehör zu verschaffen, können mit den von Plattformen und sozialen Netzwerken angebotenen Infrastrukturen und Mitteln gleichberechtigt am Informationsaustausch teilnehmen, Sichtbarkeit für ihre Anliegen erreichen und Zusammenschluss mit Gleichgesinnten finden.

Trotz dieser auf den ersten Blick erheblich erweiterten Möglichkeiten zu Informationsgewinnung, Teilhabe an Kommunikations- und Entscheidungsprozessen und sozialer Vernetzung – auch mit Menschen, die anderer Auffassung sind – werden diese Potenziale unter den aktuellen Bedingungen häufig nicht realisiert oder verdeckt von Prozessen, die menschliche Handlungsspielräume und Möglichkeiten für den öffentlichen Vernunftgebrauch und demokratische Verständigungsprozesse eher vermindern. So werden beispielsweise viele Prozesse der Onlinekommunikation durch die beschriebenen algorithmischen Formatierungen immer wieder eingeschränkt und verzerrt.

Die auf eine Maximierung der Aufmerksamkeit und Verweildauer ausgerichtete Kuratierung erhöhen so zwar einerseits den Komfort für Nutzende, wenn diese automatisch zu ihren Präferenzen passende und unterhaltsame bzw. spannende Inhalte erhalten oder in Onlinesuchen schnell und leicht die gewünschten Informationen finden. Andererseits laufen aber

gerade die nach aktuellen Kuratierungskriterien geförderten schnelllebigen und verbreitungsstarken, polarisierenden, emotional und moralisch aufgeladenen Inhalte und Reaktionen in vielerlei Hinsicht den Grundgedanken demokratischer Kommunikation und Willensbildung zuwider. Insbesondere eine stark von negativen Emotionen geprägte Polarisierung vermindert die Handlungsspielräume für den freien und sachlichen Austausch von Gründen und Argumenten. Wo alternativ oder zusätzlich Effekte wie Filterblasen und Echokammern hinzukommen – Phänomene, deren Verbreitung nach wie vor unklar ist –, vermindern sich die Möglichkeiten für konstruktive Auseinandersetzungen gegebenenfalls noch zusätzlich, da man dann konträren Auffassungen und Argumenten im Rahmen eines personalisierten Feeds mitunter kaum noch begegnet.

Der Einsatz von persuasiver personalisierter politischer Werbung, oft jenseits klar kenntlich gemachter Kampagnen und unter Einsatz von Fake-Accounts und Bots, gefährdet demokratische Prozesse nicht nur, indem sie den individuellen Vernunftgebrauch mit manipulativen Techniken erschwert, sondern auch, indem durch den täuschenden und verzerrenden Charakter solcher Aktionen das Vertrauen in den demokratischen Prozess selbst untergraben wird. Derartige Vertrauensverluste wiederum begünstigen gemeinsam mit der bereits erwähnten algorithmischen Favorisierung aufmerksamkeitserregender und negativer emotional wie moralisch aufgeladener Inhalte Verschwörungsmythen und andere kollektive Erregungen, die wiederum den öffentlichen Vernunftgebrauch erschweren, der für eine funktionierende Demokratie unerlässlich ist.

Hinzu kommt die problematische Konzentration digitaler Kommunikationsinfrastrukturen in den Händen weniger Konzerne mit vielfach als unzureichend betrachteten Verantwortlichkeiten, Rechenschaftspflichten und öffentlichen Kontrollmöglichkeiten ebenso wie die aktuell zu beobachtenden

Defizite in der Moderation problematischer Inhalte.[331] Letztere betreffen sowohl die technische Umsetzung als auch die Frage, wie eine angemessene Balance zwischen zu wenig und zu viel Kontrolle in das digitale Kommunikationsgeschehen gelingen kann. Die aktuelle Aufstellung, in der die auf kommerziellen Plattformen zum Einsatz kommenden soziotechnischen Mechanismen für die Öffentlichkeit weitgehend opak bleiben und selbst Forschung keinen Zugang zu den proprietären Algorithmen hat, geht für die Gesellschaft mit stark verminderten Handlungs- sowie Kontrollmöglichkeiten einher.

In diesem Rahmen stellt sich auch die Frage, ob eine digitale Kommunikationsinfrastruktur in öffentlicher Verantwortung Lösungsansätze für diese Probleme bieten könnte. Das Ziel wäre nicht nur, Menschen eine Alternative zu kommerziellen Plattformen anzubieten, sondern darüber hinaus durch Meinungsvielfalt, kommunikatives Ethos, Minderheitenschutz und Seriosität auch die privaten Anbieter zu einem höheren Maß an Demokratieverträglichkeit zu veranlassen. Der Demokratieentwicklung unter den Bedingungen einer dynamischen digitalen Transformation käme das zugute.

Onlineplattformen und soziale Medien bilden die zentralen Infrastrukturen für Information und Kommunikation und sind somit von entscheidender Bedeutung für die öffentliche Meinungsbildung. Gleichzeitig liegen sie jedoch in der Hand weniger globaler Akteure, die primär ökonomische Zwecke verfolgen, wodurch deren soziotechnische Praktiken der Selektion und Moderation einerseits erheblichen Einfluss auf Prozesse der Information, Kommunikation und der öffentlichen Meinungsbildung haben, andererseits sich jedoch effektiver Kontrolle der Rechtmäßigkeit häufig entziehen.

331 Diesen Problemen soll – und wird teilweise auch schon – mit europäischen und nationalen Gesetzen entgegengewirkt werden (z. B. Netzwerkdurchsetzungsgesetz, Digital Services Act, Digital Markets Act). Auch der Medienstaatsvertrag sieht solche Regelungen vor.

Empfehlungen

- *Empfehlung Kommunikation 1:* Regulierung sozialer Medien: Es bedarf klarer rechtlicher Vorgaben, in welcher Form und in welchem Ausmaß soziale Medien und Plattformen über ihre Funktions- und Vorgehensweisen zur Kuratierung und Moderation von Inhalten informieren müssen und wie dies auf der Grundlage institutioneller Regelungen umgesetzt wird. Dies muss durch externe Kontrollen überprüfbar sein; rein freiwillige Ansätze privater Handelnder, insbesondere die unverbindliche Überprüfung durch von diesen selbst besetzten Aufsichtsgremien, sind nicht ausreichend. Hier gibt es auf Ebene der Europäischen Union im Digital Services Act bereits Ansätze, die aber noch nicht weit genug gehen.
- *Empfehlung Kommunikation 2:* Transparenz über Moderations- und Kuratierungspraktiken: Anstelle allgemeiner Moderations- und Löschungsrichtlinien und wenig aussagekräftiger Zahlen über Löschungen muss für externe Kontrollen nachvollziehbar sein, wie, unter welchen Umständen und anhand welcher Kriterien solche Entscheidungen gefällt und umgesetzt werden und welche Rolle hierbei Algorithmen bzw. menschliche Moderierende übernommen haben. Darüber hinaus müssen auch die grundlegenden Funktionsweisen der Kuratierung von Inhalten sozialer Medien und Plattformen in dem Ausmaß offengelegt werden, das nötig ist, um systemische Verzerrungen und möglicherweise resultierende informationelle Dysfunktionen erkennen zu können. Die Berichtspflichten und Transparenzvorgaben im Medienstaatsvertrag, im Netzwerkdurchsetzungsgesetz und im Digital Services Act stellen dies noch nicht hinreichend sicher. Die datenschutzrechtlichen Auskunftspflichten gemäß Art. 12 ff. DSGVO sind zum Teil auf nationalstaatliche Ebene beschränkt worden und erfassen oftmals diese weiter gehenden Aspekte nicht.
- *Empfehlung Kommunikation 3:* Zugriff auf wissenschaftsrelevante Daten von Plattformen: Um die Wirkungsweisen

von Plattformen und sozialen Medien, ihren Einfluss auf öffentliche Diskurse, aber auch weitere Themen von hoher gesellschaftlicher Relevanz zu untersuchen, sollte sichergestellt werden, dass unabhängigen Forschenden der Zugriff auf wissenschaftsrelevante Daten von Plattformen nicht mit dem pauschalen Verweis auf Betriebs- oder Geschäftsgeheimnisse verweigert werden kann. Für den Zugang müssen sichere, datenschutzkonforme sowie forschungsethisch integre Wege gefunden werden. Netzwerkdurchsetzungsgesetz und Digital Services Act enthalten bereits Reglungen zum Datenzugang, die aber in ihrem Anwendungsbereich sehr begrenzt sind; auch der Data Act sieht vergleichbare Regelungen vor.

» *Empfehlung Kommunikation 4:* Berücksichtigung von Sicherheit, Datenschutz und Geheimhaltungsinteressen: Anforderungen an Offenlegungen und Datenzugang müssen kontextsensitiv spezifiziert werden, wobei Anforderungen an Sicherheit und Schutz vor Missbrauch, Datenschutz sowie dem Schutz von intellektuellem Eigentum und Geschäftsgeheimnissen angemessen Rechnung zu tragen ist. Je nach Kontext muss zwischen unterschiedlich klar definierten Zeitpunkten der Prüfung und Graden der Offenlegung unterschieden werden.

» *Empfehlung Kommunikation 5:* Personalisierte Werbung, Profiling und Microtargeting: Personalisierte Werbung ist das zentrale Geschäftsmodell sozialer Medien und Plattformen. Die Praktiken des Profilings und Microtargetings können jedoch problematische Auswirkungen auf öffentliche Kommunikation und Meinungsbildung entfalten, insbesondere im Kontext politischer Werbung. Um solche negativen Auswirkungen durch effektive Regelungen zu verhindern, ist es zunächst notwendig, die Bedingungen für eine Erforschung und Überprüfung der Zusammenhänge zwischen Geschäftsmodellen und Praktiken algorithmischer Kuratierung in ihren Wirkungsweisen und Effekten zu schaffen. Der auf Ebene der Europäischen Union

diskutierte Vorschlag für eine Verordnung über die Transparenz und das Targeting politischer Werbung adressiert diesen Bedarf. Hierbei zeigen sich allerdings auch die Herausforderungen, Regeln so zuzuschneiden, dass sie einerseits wirksam sind, andererseits aber die Freiheit der politischen Kommunikation nicht übermäßig beschränken.

» *Empfehlung Kommunikation 6:* Bessere Regulierung von Onlinemarketing und Datenhandel: Viele informationelle und kommunikative Dysfunktionen haben ihre Ursache im Onlinemarketing, welches das grundlegende Geschäftsmodell vieler sozialer Medien und Plattformen ist und auf der Sammlung, Analyse und dem Verkauf vielfältiger Daten über die Personen, die diese Angebote nutzen, beruht. Das Problem ist hierbei nicht die Werbefinanzierung per se, sondern der invasive Umgang mit diesen Daten. Hier gilt es einerseits, die Auswirkungen dieses Geschäftsmodells auf öffentliche Diskurse besser zu erforschen. Andererseits bedarf es besserer gesetzlicher Regelungen, um sowohl Individuen in ihren Grundrechten online effektiver zu schützen als auch negative systemische Effekte auf den öffentlichen Diskurs zu minimieren. In diese Richtung gehende Vorschläge hat der Deutsche Ethikrat 2017 unter dem Stichwort „Datensouveränität" in seiner Stellungnahme zu Big Data und Gesundheit vorgestellt. Europäische Regelungen wie der Digital Markets Act adressieren das Problem der Datenmacht großer Plattformen, aber – schon aus Gründen der Regelungskompetenz – nicht mit Blick auf die Folgen für den öffentlichen Diskurs.

» *Empfehlung Kommunikation 7:* Machtbeschränkung und Kontrolle: Unternehmen, die im Bereich der öffentlichen Vorstellung von Daten bzw. Tatsachen de facto monopolartige Machtmöglichkeiten haben, sind durch rechtliche Vorgaben und entsprechende Kontrolle auf Pluralismus, Minderheiten- und Diskriminierungsschutz zu verpflichten. Ein Teil der Mitglieder des Deutschen Ethikrates ist der Auffassung, dass medienrechtliche Regelungen zur

Sicherung von Pluralität, Neutralität und Objektivität generell auf Nachrichtenfunktionen von sozialen Medien und Plattformen ausgedehnt werden sollten, sofern sie denen traditioneller Medien ähneln.

- *Empfehlung Kommunikation 8:* Erweiterung der Nutzerautonomie: Plattformen und soziale Medien sollten ihre Inhalte auch ohne eine personalisierte Kuratierung verfügbar machen. Darüber hinaus sollten sie für die Kriterien, nach denen Inhalte auf Plattformen und in sozialen Medien algorithmisch ausgewählt und prioritär präsentiert werden, weitere Wahlmöglichkeiten anbieten. Dazu sollte auch die Möglichkeit gehören, bewusst Gegenpositionen angezeigt zu bekommen, die den bisher geäußerten eigenen Präferenzen zuwiderlaufen. Solche Wahlmöglichkeiten sollten gut sichtbar und leicht zugänglich sein.
- *Empfehlung Kommunikation 9:* Förderung kritischer Rezeption von Inhalten: Zur Eindämmung unreflektierter Verbreitung fragwürdiger Inhalte sollten diverse Hinweisfunktionen entwickelt und eingesetzt werden, die eine kritische Auseinandersetzung mit Material fördern, bevor man sich dafür entscheidet, es zu teilen oder öffentlich darauf zu reagieren. Dies könnten etwa Rückfragen sein, ob Texte gelesen und Videos geschaut wurden, bevor man sie teilt, oder Angaben zur Seriosität von Quellen.
- *Empfehlung Kommunikation 10:* Alternative Informations- und Kommunikationsinfrastruktur: Zu erwägen wäre, den privaten Social-Media-Angeboten im europäischen Rahmen eine digitale Kommunikationsinfrastruktur in öffentlich-rechtlicher Verantwortung zur Seite zu stellen, deren Betrieb sich nicht am Unternehmensinteresse eines möglichst langen Verweilens von Menschen auf der Plattform oder an anderen kommerziellen Interessen orientiert. Damit sollte nicht etwa der öffentlich-rechtliche Rundfunk (TV und Radio) auf eine weitere digitale Plattform ausgedehnt, sondern eine digitale Infrastruktur bereitgestellt werden, die eine Alternative zu den kommerzbetriebenen,

stark oligopolartigen Angeboten bietet. Um eine hinreichende Staatsferne zu garantieren, könnte auch an eine Trägerschaft in Gestalt einer öffentlichen Stiftung gedacht werden.

8 ÖFFENTLICHE VERWALTUNG

8.1 Einleitung

Die öffentliche Verwaltung stellt die exekutive Ebene der Umsetzung der in demokratischen Meinungsbildungs- und Entscheidungsverfahren gefassten Beschlüsse dar. Sie versteht sich als *vollziehende Gewalt des Staates*[332] und lässt sich systematisieren in die Ordnungsverwaltung (Vollzug und Kontrolle von Gesetzen), die Dienstleistungsverwaltung (Umsetzung der gesetzlich begründeten, technischen oder persönlichen Dienstleistungsansprüche aller Bürgerinnen und Bürger), die Wirtschaftsverwaltung (Bewirtschaftung aller Einnahmen, Ausgaben und Vermögen der öffentlichen Hand), die Organisationsverwaltung (Personalwesen einschließlich Personalrekrutierung, Fort- und Weiterbildung) und die politische Verwaltung (insbesondere Zuarbeit für die politische Führung sowie Entscheidungstragende in der Legislative).[333]

Für Menschen, aber auch viele Organisationen stellt die öffentliche Verwaltung, so etwa im Finanz-, Steuer-, Melde- und Sozialwesen oder in der Straffälligen- und Jugendgerichtshilfe, die unmittelbar erfahrbare Staatsgewalt dar. Ihre Exekutivfunktion darf dabei nicht als bloß ausführende Vollstreckung bereits anderweitig getroffener Entscheidungen verstanden werden. Zwar müssen übergeordnete und direkt demokratisch legitimierte Entscheidungen, etwa im Sozialrecht oder Steuerrecht, eingehalten und umgesetzt werden; dennoch verbleiben auf der Ebene der Verwaltung weitreichende Beurteilungs- und Entscheidungsspielräume in der einzelfallbezogenen Adaptation allgemein formulierter Gesetze und Verordnungen, so zum Beispiel unter dem Begriff der Verhältnismäßigkeit, bei unbestimmten Rechtsbegriffen oder Risikoentscheidungen.

332 Schmidt-Aßmann 1998, 148 ff.
333 Schmidt 2010, 859 f.; Hesse/Ellwein, 308 ff.

Zudem sind der öffentlichen Verwaltung von Gesetzes wegen mitunter unmittelbare Entscheidungsbefugnisse zugewiesen – auch in der Dienstleistungsverwaltung etwa des Gesundheits- und Sozialwesens. Gerade hier unterliegen die Mitarbeitenden der öffentlichen Verwaltung einer spezifischen Professionsethik.[334] Funktionierende, transparente, als legitim anerkannte und bürgernahe Verwaltung ist daher für ein funktionierendes Gemeinwesen und die Akzeptanz von Demokratie und Staat wesentlich.

8.2 Ethische Fragen algorithmischer Automatisierung im Verwaltungshandeln

Seit den 1970er-Jahren werden Konzepte für eine gleichzeitig effizientere wie auch bürgernähere öffentliche Verwaltung im Rahmen von Digitalisierungsstrategien entwickelt, erprobt und teils umgesetzt.[335] Damit verbinden sich unter anderem Hoffnungen auf eine Rationalisierung und Beschleunigung staatlichen Verwaltungshandelns, eine effektivere und kohärentere Datennutzung sowie eine Ausweitung der Einbeziehung wissenschaftlichen und bürgerschaftlichen Sachverstandes. Dem steht die Schreckensvision einer sogenannten „Algokratie" gegenüber, in der autonome Softwaresysteme die staatliche Herrschaft über Menschen ausüben, Bürgerinnen und Bürgern durchgehend Entscheidungen unterworfen sind, deren Algorithmen intransparent sind und keinen Widerspruch dulden.[336] Werden in den positiven Stimmen die Potenziale der Bürgernähe und der niedrigschwelligen digitalen Erreichbarkeit von Verwaltungsleistungen vom Wohnzimmer

334 Lob-Hüdepohl 2002; Trappe 2013.
335 Simitis/Hornung/Spiecker gen. Döhmann 2019, Einleitung Rn. 6; zu den veränderten Vorstellungen von Verwaltung in diesem Zuge Rn. 7.
336 Der Begriff der Algokratie geht zurück auf Aneesh 2009; vgl. auch Danaher 2016.

aus betont, so sehen kritische Stimmen digitale Technologien als weiteren Schritt zu einer technokratischen Bürokratie, in der Kommunikation hinter anonymen Datenmengen und standardisierten, noch dazu schwer verständlichen Bedienoberflächen verschwindet.

Tatsächlich lässt sich in den vergangenen Jahren in vielen Ländern in und außerhalb Europas ein zunehmender Einsatz von automatisierten Entscheidungssystemen in der öffentlichen Verwaltung beobachten. Beispiele reichen von der Bewertung von Arbeitsmarktchancen Jobsuchender in Österreich[337] und Polen[338] über die Verwendung von Software für die Prüfung und Vergabe von Sozialleistungen in England[339], Frankreich[340] und den Niederlanden[341] bis hin zu prädiktiven Analysen im Betreuungs- und Fürsorgebereich in Finnland[342] und Spanien[343], aber auch im Bereich der Polizei. Auch wenn in Deutschland der Einsatz von automatisierten Entscheidungssystemen in der öffentlichen Verwaltung noch selten ist[344], finden sich auch hier einzelne Projekte wie etwa die automatische Berechnung des Arbeitslosengelds durch das IT-System ALLEGRO der Bundesagentur für Arbeit[345] oder die in Hamburg zum Einsatz kommende Fachanwendung JUS-IT für die Koordination und Abrechnung von Sozialdiensten[346].

Vielen Vorteilen solcher Systeme, wie der Steigerung der Effizienz von Verwaltungsvorgängen und der besseren Absicherung von Entscheidungen angesichts häufig komplexer

337 https://www.derstandard.at/story/2000122684131/gericht-macht-weg-fuer-umstrittenen-ams-algorithmus-frei [22.02.2023].
338 Panoptykon Foundation 2015.
339 https://www.theguardian.com/technology/2019/oct/14/fears-rise-in-benefits-system-automation-could-plunge-claimants-deeper-into-poverty [22.02.2023].
340 https://algorithmwatch.org/en/robo-debt-france [22.02.2023].
341 https://algorithmwatch.org/de/risikobuerger [22.02.2023].
342 AlgorithmWatch 2020, 84 ff.
343 AlgorithmWatch 2019, 117 ff.
344 https://netzpolitik.org/2020/automating-society-report-2020-automatisierung-schreitet-auch-in-deutschland-voran [22.02.2023].
345 https://algorithmwatch.org/de/hartz-iv-algorithmen-diskriminierung [30.01.2023].
346 https://www.hamburg.de/jus-it [30.01.2023].

Datenlagen, stehen Risiken an anderen Stellen gegenüber. Hier ist zunächst einmal die Qualität der verwendeten Systeme zu bewerten und die Frage zu stellen, ob und in welchem Umfang die verwendeten Systeme Diagnosen und Prognosen tatsächlich verbessern. Es stellt sich in diesem Kontext auch die Frage, ob die Genauigkeit für verschiedene Anwendungskontexte oder Personengruppen gleich ist oder ob es möglicherweise systematische Verzerrungen oder Diskriminierungen gibt (Algorithmic Bias). In ethischer Hinsicht sind hierbei insbesondere zwei Themenkomplexe von besonderer Bedeutung, die einerseits Fragen von Autonomie, Autorschaft und Verantwortung sowie andererseits Fragen der Gerechtigkeit berühren.

Im ersten Fall geht es um die Frage, ob und wie die Nutzung von KI zur Unterstützung von Entscheidungen oder gar das vollständige Delegieren von Entscheidungen an KI-Systeme menschliche Handlungsfähigkeiten und Autorschaft beeinflusst. Bereits in den Bezeichnungen für solche Systeme wird diese Problematik deutlich. Der häufig verwendete Begriff „ADM-System" wird teils in *algorithmic decision making*, teils in *automated decision making* aufgelöst. In beiden Fällen wird nahegelegt, dass die Entscheidungen vollständig automatisiert erfolgen oder aber Entscheidungen vollständig an Maschinen delegiert werden. Demgegenüber steht beispielsweise der Begriff „Decision Support System", der nahelegt, dass KI-Systeme menschliche Entscheidungen lediglich unterstützen sollen. Diese Unterscheidung kommt auch in der Datenschutz-Grundverordnung zum Tragen, wenn in Artikel 22 postuliert wird, dass niemand zum Objekt einer allein auf Algorithmen basierenden Bewertung gemacht werden sollte, sondern belastende Wertungsentscheidungen von einem Menschen verantwortet werden sollten. Mit Blick auf den Automation Bias, das heißt die menschliche Tendenz, sich maschinellen Empfehlungen vorbehaltslos anzuschließen, stellt sich allerdings die Frage, ob dieses Problem nur bei Entscheidungen ohne jedes menschliche Eingreifen und nicht bereits im Falle der Entscheidungsunterstützung relevant wird.

Schon bei der Nutzung von Software zur Entscheidungsunterstützung, beispielsweise mithilfe von Risikoscores, ändert sich die Rolle der menschlichen Verantwortungs- und Entscheidungsträger grundlegend. Nach einer Phase der Bewährung dieser Systeme kann es leicht zur Dominanz einer Default-Praxis des Verwaltungshandelns in diesem Bereich kommen: Es treten Gewöhnung und Routine ein. Die maschinelle Interpretation der Daten gibt im Regelfall eine Entscheidung vor, von der nur in besonderen Einzelfällen, für die es dann Gründe geben muss, abgewichen wird. Die Entscheidung gegen eine maschinell vorbereitete Empfehlung, wenn der Softwareeinsatz etabliert und im Ganzen bewährt ist, legt der verantwortlichen Person Recherche- und Begründungspflichten auf, die bei komplexen Sachverhalten einen erheblichen Aufwand nach sich ziehen. Daher können auch bei formaler Beschränkung von ADM-Systemen auf Entscheidungsvorbereitung und -unterstützung die Verantwortungsspielräume von Verwaltungspersonal de facto eingeschränkt werden. Umgekehrt und angesichts der Tatsache, dass auch das traditionelle Verwaltungshandeln von Routinen und Mustern geprägt ist, was im Einzelfall immer wieder zu Fehlentscheidungen führt und der notwendigen Flexibilität bei komplexeren Sachverhalten entgegensteht, kann die Ausweitung der Entscheidungsunterstützung durch ADM-Systeme jedoch auch helfen, starre Routinen aufzubrechen, komplexere Entscheidungssituationen rechtzeitig zu diagnostizieren und eine angemessene Reaktion erst zu ermöglichen.

Wenn Entscheidungen von Softwaresystemen unterstützt oder gar vollständig an diese delegiert werden, stellen sich auch Fragen der Verantwortungszuschreibung. Auch beim vollautomatisierten Verwaltungshandeln muss das Recht der Bürgerschaft gewahrt werden, sich mit den in der Verwaltung Verantwortlichen auseinanderzusetzen, Einspruch zu erheben und letztlich, wenn keine Einigung erzielt werden kann, zu versuchen, ihr Recht vor dem Verwaltungsgericht durchzusetzen. Daher muss es eine klare und für sie erkennbare

Verantwortungszuweisung an eine dafür zuständige und somit verantwortliche Behörde geben.

Das zweite Problem betrifft Fragen der Gerechtigkeit und der Verhinderung von Diskriminierung. Auch hier wird der Begriff des Bias verwendet, jedoch mit einer völlig anderen Bedeutung: Im Gegensatz zum zuvor geschilderten Fall des Automation Bias geht es hier nicht um kognitive Verzerrungen menschlicher Akteure, sondern um systematische Verzerrungen im Output der KI-Systeme, die zuweilen als Technical Bias[347] oder Algorithmic Bias beschrieben werden. Ein viel diskutiertes Beispiel hierfür ist die Software COMPAS, die in den USA zur Ermittlung von Risikoprofilen für ehemalige straffällige Personen eingesetzt wird. Hier konnte nachgewiesen werden, dass die Prognosen insofern diskriminierend waren, als dunkelhäutige Menschen deutlich häufiger eine fälschlicherweise zu negative Prognose erhielten als hellhäutige Menschen, die im Gegensatz dazu fälschlicherweise zu positive Prognosen bekamen. Häufige Ursache solch diskriminierender Systeme sind Verzerrungen in den Trainingsdaten, aber auch verschiedene methodische Entscheidungen[348]. Das Resultat ist, dass

347 Friedman/Nissenbaum 1996.

348 Barocas/Selbst (2016) liefern hier eine ausgesprochen hilfreiche Übersicht möglicher Quellen von Bias und Diskriminierung in datenbasierten Systemen. Neben der Auswahl von Zielvariablen, Labels oder relevanten Merkmalen sind hier vor allem die verwendeten Daten von entscheidender Bedeutung. Hierbei kann es nicht nur durch Fehlerhaftigkeit der Daten selbst, sondern auch durch Über- und Unterrepräsentativität zu Problemen kommen, das heißt, wenn zum Beispiel Personen in Datensätzen überproportional oder unterproportional repräsentiert sind. Darüber hinaus besteht das Problem sogenannter Proxyvariablen: Sensible Kategorien wie beispielsweise Geschlecht, religiöse Zugehörigkeit oder sexuelle Orientierung lassen sich teilweise aus anderen Datenpunkten wie etwa Hobbies, Einkaufsverhalten oder Wohnort erschließen. Aufgrund dieser sogenannten redundanten Encodierung können Personen aufgrund ihres Geschlechts oder ihrer sexuellen oder religiösen Orientierung durch Software auch dann systematisch diskriminiert werden, wenn diese Angaben gar nicht erhoben wurden – eben weil diese Kategorien aus anderen Daten ableitbar sind. Die meisten der zuvor geschilderten Quellen von Bias und Diskriminierung sind nicht beabsichtigt, sondern das Resultat methodischer Entscheidungen bzw. Fehler. Es ist jedoch prinzipiell auch möglich, absichtlich diskriminierende Effekt in Softwaresysteme einzubauen – ein Problem, welches Barocas/Selbst Verschleierung (*masking*) nennen.

insbesondere datenbasierte Systeme Stereotypen, aber auch gesellschaftliche Ungleichheit in scheinbar neutrale Systeme übersetzen und verschleiern und somit existierende soziale Ungleichheit weiter in die Zukunft fortschreiben. Dies ist jedoch keine zwangsläufige Entwicklung. Im Gegenteil können datenbasierte Systeme solche historischen Ungerechtigkeiten auch zuvorderst aufdecken und sie damit für Gegenmaßnahmen zugänglich machen.

Eine grundsätzliche Grenze für die Anwendung von ADM-Systemen liegt in nicht eliminierbaren normativen Ziel- oder Regelkonflikten im deutschen, deontologisch verfassten Rechtssystem. Folgenabwägung allein bestimmt in der deontologisch orientierten Ordnung in Deutschland nicht das Rechtmäßige. Sie kommt zwar zum Einsatz, zum Beispiel im Zusammenhang mit der Verhältnismäßigkeit, wenn es darum geht, das mildeste Mittel im Sinne eines Eingriffs in Grundrechte zu bestimmen, das dennoch für die Zielerreichung effektiv ist. Aber in der Bestimmung des deontologisch Rechtmäßigen geht es nicht um das mehr oder weniger Vorteilhafte und nicht um Nutzenoptimierung in Bezug auf die Folgen, sondern um die Frage, wer welche Rechte hat und wie diese im Konfliktfall so abzuwägen und zu gewichten sind, dass unbedingte Ansprüche auf Schutz der Person gewahrt und ihre Verrechenbarkeit (im Sinne einer reinen Nutzenoptimierung) verhindert werden kann.[349] Insoweit sind der quantitativen Berechnung Grenzen gesetzt; vielmehr es geht um eine nur menschlicher Urteilskraft zugängliche Bewertung. So müssen in menschlicher Deliberation Argumente und Gegenargumente vorgebracht werden, auf deren Grundlage am Ende entschieden wird (vgl. Abschnitt 3.2.2). Hier zeigt sich eine ultimative Grenze der Algorithmisierung ethischer und rechtlicher Entscheidungsprozesse.

Diese grundsätzlichen Überlegungen sprechen nicht dagegen, Teilbereiche der Entscheidungsprozesse zu algorithmisieren und die rechtliche und moralische Deliberation mit

349 Ross 1930.

Softwaresystemen zu unterstützen. Im Gegenteil geht es darum, auch im Verwaltungshandeln die Vorteile von digitalen und insbesondere KI-gestützten Systemen zu nutzen, ohne jedoch in die genannten technokratischen Fallen zu tappen. Die Gestaltung entsprechender Systeme kann proaktiv von Beginn an so angelegt werden, dass ethische Werte und Normen berücksichtigt werden, wie dies beispielsweise im Ansatz des Value Sensitive Design angestrebt wird.[350] Dies kann der Unterstützung und Erweiterung menschlicher Autorschaft dienen.

Im Folgenden werden diese Fragen anhand von Beispielen aus den Bereichen des Sozialwesens (vgl. Abschnitt 8.3) und des Polizeiwesens (vgl. Abschnitt 8.4) exemplarisch illustriert und vertieft. Die beiden Felder wurden ausgewählt, weil in ihnen häufig Entscheidungen von großer Tragweite und möglicherweise großer Eingriffstiefe vorgenommen werden. Daher erscheint es einerseits geboten, Technologien zu nutzen, welche die Qualität dieser Entscheidungen verbessern können, indem sie beispielsweise genauere Vorhersagen liefern und Fehlentscheidungen verringern. Ob KI-Technologien die Entscheidungsqualität jedoch verbessern, hängt von deren Qualität ab. So gilt es einerseits zu prüfen, ob und in welchem Ausmaß es durch den Einsatz von Software zur Entscheidungsunterstützung tatsächlich zu weniger Fehlern in beide Richtungen, das heißt weniger falsch-positiven und weniger falsch-negativen Resultaten kommt. Zum anderen ist zu prüfen, ob sich Genauigkeit und Fehleranfälligkeit für unterschiedliche Personengruppen unterscheiden, um mögliche diskriminierende Effekte zu vermeiden.

350 Friedman/Hendry 2019.

8.3 Automatische Entscheidungssysteme am Beispiel des Sozialwesens

Im Sozialwesen sind auf der Grundlage einschlägiger (sozial-) rechtlicher Gesetze bzw. Ausführungsbestimmungen Entscheidungen mit weitreichenden Folgen für die Betroffenen zu treffen, so etwa über die Gewährung von Hilfen, bei Maßnahmen im Kontext einer Kindeswohlgefährdung oder bei der Abschätzung von Gefährdungspotenzialen straffälliger Personen in der Bewährungshilfe. Solche Entscheidungen werden von den involvierten Fachkräften der Sozialverwaltungen für gewöhnlich unter enger Einbindung der Betroffenen und Leistungsberechtigten getroffen. So umfasst der Schutzauftrag des Jugendamtes bei Kindeswohlgefährdung nicht nur die Erhebung des Gefährdungsrisikos unter Einbeziehung aller relevanten Fachkräfte, die mit dem betroffenen Kind und/oder den Personensorgeberechtigten beruflich in Kontakt stehen, sondern auch die unmittelbare Einbindung der Erziehungsberechtigten sowie des betroffenen Kindes (§ 8a SGB VIII). Oftmals münden Entscheidungen der öffentlichen Verwaltung in konkrete Hilfeplanungen, die auf der Basis von Hilfeplanbesprechungen oder sogar förmlichen Hilfeplankonferenzen gemeinsam mit den Betroffenen erstellt und verbindlich vereinbart werden, nachdem ebenfalls gemeinsam auf der Basis einer Ressourcen- und Potenzialanalyse der entsprechende Hilfebedarf erhoben wurde. Im Zentrum der Entscheidungsprozesse stehen deshalb Gespräche, in denen die lebensweltliche Expertise der Betroffenen bzw. der Leistungsberechtigten mit der professionellen Expertise der Fachkräfte verknüpft werden.[351] Dies dient nicht nur der besseren Ergebnissicherung der professionellen Entscheidung, sondern unmittelbar auch der Steigerung der Lebensführungskompetenz der Betroffenen,

351 Schwabe 2008; Moch 2018. Zur (verdrängten) Fehlerhaftigkeit intuitionsgestützten Entscheidens und der dringenden Notwendigkeit verbesserter Entscheidungsgrundlagen vgl. ausführlich auch Schwabe 2022.

insofern sie durch die gemeinsame Problemdiagnose auch der eigenen Ressourcen gewahr werden können.

Die Implementierung algorithmenbasierter Entscheidungshilfen in die unterschiedlichen Arbeitsbereiche des Sozialwesens[352] steht in Deutschland erst am Anfang.[353] Andere europäische oder transatlantische Länder weisen bereits ein erheblich höheres Implementierungsniveau auf. Gleichwohl befinden sich auch in Deutschland eine Reihe von Systemen in der Einführungs-, Erprobungs- oder Vorbereitungsphase.[354] Eine ethische Einordnung, die auch die internationale Entwicklung im Blick hat, ist deshalb auch für Deutschland von großem Interesse. Diese soll in den nächsten Schritten entlang der Unterscheidung von Erweiterung und Verminderung professioneller Handlungsoptionen erfolgen, wobei es gerade im Sozialwesen von Bedeutung ist, die Erweiterungs- bzw. Verminderungsdimensionen gerade auch aus der Perspektive der von den Entscheidungen jeweils betroffenen Personen in den Blick zu nehmen.

8.3.1 Erweiterung professioneller Handlungskompetenz

Unbeschadet der fachlichen Bedeutsamkeit je spezifischer Personenorientierung im Rahmen von Entscheidungsprozessen bieten Verfahren der Entscheidungsfindung, die auf einem gewissen Maß an Standardisierung – auch und gerade unter Einbeziehung algorithmenbasierter Erfassung und Steuerung – beruhen, gegenüber herkömmlichen Verfahren wichtige Vorteile. Häufig hängen Entscheidungen über die Gewährung von Hilfen und anderen Interventionen im Sozialwesen im Wesentlichen von intuitiven Einschätzungen der Fachkräfte

352 Schneider 2020.
353 Vgl. entsprechende Beiträge in Kutscher et al. 2020 – insbesondere „Teil VI: Digitalisierung in Handlungsfeldern der Sozialen Arbeit".
354 AlgorithmWatch 2020, 114 ff.

ab, selbst da, wo eine stärkere Standardisierung von Entscheidungsfindungen möglich wäre. Solche intuitiven Einschätzungen profitieren zwar von der beruflichen Erfahrung und dem in den relevanten Bereichen des Sozialwesens durchaus stark verbreiteten innerkollegialen Austausch. Wie eine hohe Zahl empirischer Studien gleichwohl belegt, bleiben sie aber hinsichtlich ihrer Sachangemessenheit unter dem Qualitätsniveau „statistisch-aktuarischer“[355] oder sogar „mustererkennend-statistischer“[356] Ansätze mittlerweile deutlich zurück – möglicherweise zum Nachteil der von den Entscheidungen Betroffenen. Dies zeigt sich besonders bei der Generierung prognostischen Wissens, also bei der Einschätzung über Gefährdungs- oder Entwicklungspotenziale hilfebedürftiger Personen. Für soziale Professionen, bei denen eine gute Wirkung von Maßnahmen im Mittelpunkt stehen muss, dürfen die Vorteile evidenzbasierter Instrumente daher nicht vernachlässigt werden.[357] In dieser Hinsicht kann die Verwendung von Softwaresystemen zur Entscheidungsunterstützung grundsätzlich zur Erweiterung professioneller Handlungskompetenz beitragen. Verbesserte Ergebnisse professioneller Entscheidungen führen zudem in der Regel zu erweiterten Gestaltungsspielräumen der professionell unterstützten Hilfeempfangenden und dies ist der entscheidende Maßstab zur Bewertung der Güte sozialprofessioneller Interventionen.

Eine Verbesserung der Entscheidungsgrundlage ist besonders wichtig bei der Abschätzung von Gefährdungspotenzialen, denen entweder besonders vulnerable Personen (Stichwort „Kindeswohlgefährdung“, siehe Infokasten 10) ausgesetzt sind oder die etwa von straffällig gewordenen Personen für Dritte

355 Rettenberger 2018.
356 Sowohl der aktuarialistische als auch der mustererkennende Ansatz ist ein mechanisches Prognoseverfahren und beruht auf speziellen Berechnungen. Während das aktuarialistische Verfahren statistische Modelle verwendet, die stets durch Studien validiert werden müssen, sind die mustererkennenden Modelle zur Risikoeinschätzung dynamisch angelegt (Schrödter/Bastian/Taylor 2020).
357 Lob-Hüdepohl 2021.

bzw. für die Allgemeinheit insgesamt ausgehen (siehe Infokasten 11).[358] Internationale Studien belegen, dass die Treffsicherheit von Prognosen zur Eintrittswahrscheinlichkeit, die auf der Basis standardisiert-versicherungsmathematischer Berechnungen vorgenommen werden, deutlich höher ist als jene, die sich mehr oder minder allein auf die zwar fachlich geschulte, gleichwohl wesentlich intuitivere Einschätzung von Fachkräften selbst mit hoher beruflicher Erfahrung stützt.[359] Aus diesem Grund bilden besonders in den angelsächsischen Ländern prädiktive Risikomodelle mittlerweile einen festen Bestandteil in der Entscheidung von Fachkräften bei der Frage, ob in einer bestimmten Fallkonstellation die Familiensituation zum gegenwärtigen Zeitpunkt oder in Zukunft die Gefährdungslage für ein Kind ein solches Ausmaß angenommen hat oder haben wird, dass das gefährdete Kind in staatliche Obhut genommen werden sollte oder aber noch in der Familie verbleiben kann.[360]

Eine vorschnelle präventive Inobhutnahme verbietet sich nicht nur deshalb, weil damit elterliche Sorgerechte und damit Kinderrechte insgesamt berührt sind, sondern weil auch die Herausnahme eines Kindes aus der Familie etwa durch den Verlust von sozialen Alltagsbindungen sein Wohl belastet. Damit stehen sich in diesen Entscheidungskonflikten in erster Linie nicht die moralischen Ansprüche zweier unterschiedlicher Personen (Eltern gegen Kind) gegenüber, wohl aber zwei moralisch relevante Güter des Kindes (z. B. der Schutz vor übergriffigem, gewaltförmigem Verhalten der Eltern und der Schutz der gewachsenen emotionalen Bindungen innerhalb der Familie), die beide zum Kindeswohl gehören, in der konkreten Situation aber kollidieren. Bestrebungen gehen nun dahin, diese ethische Güterabwägung mithilfe algorithmenbasierter

358 Butz et al. 2021.
359 Bastian 2012; Bastian/Freres/Schrödter 2017; Bastian et al. 2018; Gillingham 2021.
360 Gillingham 2021.

Entscheidungsunterstützung auf eine bessere Grundlage zu stellen.[361]

Infokasten 10: Kindeswohlgefährdung

ADM-Systeme im Kontext der Entscheidungsfindung bei Kindeswohlgefährdung sind mittlerweile auf nahezu allen Kontinenten wenigstens in ersten Schritten eingeführt. Manche dieser Anwendungen sind hoch umstritten, etwa aufgrund von Ungenauigkeit, Fehleranfälligkeit oder diskriminierenden Effekten.

In Deutschland befindet sich die Einführung von ADM-Systemen im Bereich der Kinder- und Jugendhilfe zur Gefährdungsabschätzung des Kindeswohls in der Planungs- und Diskussionsphase. Bereits etablierter Softwareeinsatz im Bereich des Kindesschutzes schafft dazu die nötigen Voraussetzungen. OK.JUS[362], das Bestandteil einer breiten Palette von Digitalanwendungen in der öffentlichen Verwaltung ist, unterstützt Jugendämter bzw. die fallführenden Professionellen in der Wahrnehmung ihres Schutzauftrags bei Kindeswohlgefährdung gemäß § 8a SGB VIII. Dieser Schutzauftrag besteht darin, dass im Falle von gemeldeten Hinweisen auf eine Kindeswohlgefährdung das Jugendamt auf der Basis bisheriger Erkenntnisse zuverlässig und angemessen reagiert. OK.JUS dokumentiert nach vordefinierten Standards die verschiedenen Arbeitsabläufe ab dem Meldeeingang bis zu etwaigen Interventionen sowie die Bewertung von Meldungen einschließlich die zur Entscheidung notwendigen Unterlagen. Damit entstehen elektronische Akten, die in Verbindung mit OK.JUS CAP, einer Controlling- und Analyseplattform mit einer Vielzahl von Einzelberichten, der fallführenden Fachkraft eine fundiertere Einschätzung der Kindeswohlgefährdung ermöglichen. Zugleich entsteht ein umfangreicher Datenpool, auf den das ADM-System zukünftig zugreifen kann.

Deutlich weiterentwickelt sind ADM-Systeme in anderen europäischen Ländern bzw. in Übersee.[363] In Dänemark etwa war das Gladsaxe-Modell – benannt nach einem Ort im Umkreis Kopenhagens – eines der ersten Profilingsysteme, das verschiedene Risikoindikatoren kombiniert, um frühzeitig gefährdete Kinder in ihren Familien zu erkennen. Es verbindet und gewichtet elternbezogene Sachverhalte wie Erwerbslosigkeit, seelische Gesundheit, nicht eingehaltene medizinische bzw. zahnmedizinische (Termin-) Vereinbarungen oder auch Scheidungen. Die endgültige Einführung wurde nach ersten Pilotphasen auch aufgrund erheblicher Proteste zurückgestellt; das System wurde aber in Forschungskontexten weiterverwendet.[364]

In den Niederlanden starteten bereits Ende der 2000er-Jahre Versuche, das in den USA entwickelte California Family Risk Assessment für den niederländischen Kinderschutz nutzbar zu machen und auf seine prognostische

361 Vgl. etwa das Forschungsprojekt „KAIMo – Kann ein Algorithmus im Konflikt moralisch kalkulieren?“ (www.kaimo.bayern); Gutwald et al. 2021.
362 https://www.akdb.de/loesungen/oksoziales/okjus [16.01.2023].
363 Für eine vergleichende Studie vgl. Drake et al. 2020.
364 AlgorithmWatch 2020, 52 f.

Validität hinsichtlich des Risikos für Gewalt oder Vernachlässigung von Kindern in problemaffinen Familien hin zu überprüfen.[365] Dieses algorithmenbasierte Assessment-Tool zielt auf die Stärkung standardisierter bzw. strukturierter Fallführung (*structured decision making*). Es kombiniert zehn verschiedene Items, die das Risiko einer zukünftigen Vernachlässigung indizieren, mit weiteren zehn, die das Risiko eines zukünftigen Missbrauchs anzeigen sollen. Neben Aspekten wie unsichere Wohnverhältnisse oder früher eingetretene Gefährdungssituationen beziehen sich viele Items auf die (elterlichen) Sorgeberechtigten (z. B. seelische Gesundheit, Suchtprobleme, Rechtfertigung von Missbrauch durch den Sorgeberechtigten, Dominanz und Strenge). Studien ergaben eine beachtliche Treffsicherheit dieses Prognoseinstruments für ein frühzeitiges Erkennen von Gefährdungslagen.[366]

Demgegenüber legen Studien aus Neuseeland und den USA nahe, dass sich die Erwartungen an ADM-Systeme mit Blick auf eine verbesserte prädiktive Analytik nicht im gewünschten Umfang erfüllen.[367] Das Predictive Risk Modelling des neuseeländischen Centre for Applied Research in Economics und das Allegheny Family Screening Tool in Pennsylvania (USA) erweisen sich etwa aufgrund der Fehlerhaftigkeit der genutzten bzw. algorithmisch verarbeiteten Items als zu wenig präzise oder sogar als offensichtlich verfälschend und diskriminierend, was die Spielräume für die Lebensgestaltung der Betroffenen erheblich vermindert.

Eine ähnliche prädiktive Risikodiagnostik erfolgt auch in der Bewährungshilfe.[368] Dort sind Fachkräfte gehalten, Art und Weise ihrer Bewährungsführung dem Gefährdungspotenzial, das von der unter Bewährungsaufsicht stehenden Person nach ihrer Haftverbüßung für andere ausgeht, anzupassen und entsprechend engmaschig oder weitläufig anzulegen. Auch hier ist die Entscheidung keinesfalls trivial: Eine engmaschige, also mit vielen kontrollierenden Auflagen und Maßnahmen belegte Bewährungsführung geht automatisch zulasten der Freiheit der bewährungspflichtigen Person. Umgekehrt könnten bei einer weitläufigen Bewährungsführung unzumutbare Risiken

365 van der Put et al. 2016.
366 Ebd.
367 Gillingham 2016; 2021.
368 In ähnlicher Weise gehen Einschätzungen über die Legalprognose bei richterlichen Entscheidungen über das Aussetzen einer Freiheitsstrafe gemäß §§ 56, 57, 57a StGB ein. Auch hier eröffnet sich wie bei der Strafzumessung insgesamt ein breites Spektrum denkbarer Einsatzfelder (Kaspar/Höffler/Harrendorf 2020). Diese Einsatzmöglichkeiten werden im Folgenden nicht weiter thematisiert, da sie nicht in den Bereich der öffentlichen Verwaltung fallen.

für Dritte ausgehen. Deshalb versucht etwa der in den USA entwickelte und zwischenzeitlich auch in Deutschland in einigen Bundesländern etablierte Risk-Need-Responsivity-Ansatz[369] mithilfe von algorithmenbasierten statistisch-aktuarischen Prognoseinstrumenten, das individuelle Rückfallrisiko mehr oder minder valide zu ermitteln[370] und damit – wie bei Entscheidungen im Rahmen der Kindeswohlgefährdung – gerichtsfest zu machen.

Infokasten 11: Bewährungshilfe

In Deutschland orientiert sich die Bewährungshilfe in der Abschätzung von Gefährdungsrisiken überwiegend am Risk-Need-Responsivity-Ansatz. Dieser bemisst die Leistungsgestaltung der Bewährungshilfe am jeweiligen Gefährdungsrisiko, den daraus sich ableitenden Bedarfen sowie der spezifischen Ansprechbarkeit der leistungsempfangenden Person. Bei der individuellen Rückfallrisikoabschätzung werden acht Risikofaktoren (kriminelle Vorbelastung, prokriminelle Einstellungen, prokriminelle Kontakte, antisoziale/dissoziale Persönlichkeitsbezüge, Bindungen im Bereich Ehe/Familie, Bindungen im Bereich Schule/Arbeit, Substanzmittelmissbrauch, Freizeitaktivitäten) einbezogen.[371] Diese Risikofaktoren werden unter Zuhilfenahme von statistisch-aktuarischen Prognoseinstrumenten gewichtet. Sie bilden die Grundlage für die von der jeweils zuständigen Fachkraft durchzuführende Risikoabschätzung, die in den einzelnen Bundesländern allerdings unterschiedlich strukturiert und teilweise standardisiert ist.[372]

In der Schweiz sind ADM-Systeme im Rahmen des Risikoorientierten Sanktionenvollzugs (ROS) in Vollzugs- wie Bewährungsdiensten bereits etabliert.[373] ROS sieht die Abklärung des individuellen Rückfallrisikos einer gewalttätigen Person im Rahmen eines vierstufigen Prozesses vor, der erstens aus einer Triage, zweitens aus einer genaueren Abklärung des Falles, drittens aus der Planung etwaiger Interventionsinstrumente und viertens aus dem kontinuierlich ausgewerteten Verlauf besteht. Kern der ersten Stufe, in der die Dringlichkeit (Triage) einer genaueren Risikoabschätzung abgeklärt wird, bildet ein Fall-Screening-Tool (FaST), das als ADM-System die jeweiligen Fälle drei Kategorien zuordnet: Die Fälle der Kategorie A mit einem geringen Rückfallrisiko benötigen keinerlei weitere Risikoprognosen; die Fälle der Kategorie B mit einem erhöhten Rückfallrisiko benötigen als weiteren Risikoprognoseschritt eine Kurzabklärung, die die fallführende Fachkraft mittels einer Checkliste aus verschiedenen Quellen (z. B. Strafregister, Gutachten) erstellt (Fallresümee); die Fälle der Kategorie C mit einem hohen

369 Polaschek 2012.
370 Cornel/Pruin 2021.
371 Andrews/Bonta 2010, 44 ff.
372 Cornel/Pruin 2021.
373 Treuthardt/Loewe-Baur/Kröger 2018.

Rückfallrisiko bedürfen der präzisen Beurteilung aller verfügbaren Informationen durch eine forensisch geschulte psychologische Fachkraft (Risikoabklärung) im Sinne einer strukturierten Urteilsbildung. Das FaST selbst kombiniert und gewichtet unterschiedliche Items und klassifiziert den jeweils anstehenden Fall auf der Basis statistisch-aktuarischer Verfahren.[374]

Bereits seit Anfang der 2000er-Jahre kommt in mehreren US-Bundesstaaten die aufgrund diskriminierender Prognosen sehr umstrittene Software COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) zum Einsatz.[375] Sie kann sowohl bei der richterlichen Strafzumessung wie beim nachfolgenden Strafvollzug oder beim späteren Bewährungsdienst zur Anwendung kommen. Es wurde im Bemühen einer evidenzbasierten Entscheidungsfindung eingeführt, um in den verschiedenen Phasen eines Fallverlaufes sowohl aufseiten der entscheidenden Person das Ungewissheitspotenzial als auch aufseiten der Bevölkerung das Gefährdungspotenzial durch möglicherweise rückfällige gewalttätige Personen zu verringern. COMPAS selbst kommt sowohl im Kontext der Untersuchungshaft wie auch während und nach der Haftzeit (Bewährung) zur Anwendung. Es kombiniert 137 Items und differenziert damit drei Grobklassifizierungen: niedriges, mittleres und hohes Rückfallrisiko.[376] COMPAS umfasst neben einer Risikoprognose auch ein darauf aufbauendes *needs assessment*, das in die weitere Interventionsplanung eingeht. COMPAS gilt in der internationalen Debatte über die Einführung von ADM-Systemen in Vollzugs- und Bewährungsdienste trotz – oder gerade wegen – der erheblichen Vorbehalte, die ihm national und international gegenüberstehen, als zentraler Bezugspunkt.[377]

8.3.2 Verminderung von Entscheidungskompetenz, Handlungsoptionen und Autorschaft

Nachdem vorausgehend insbesondere die Erweiterung von Handlungskompetenz im Mittelpunkt stand, ist demgegenüber zu konstatieren, dass menschliche Autorschaft unter Zuhilfenahme sachdienlicher Ergebnisse von KI-Algorithmen allerdings auch vermindert werden kann. Auf der professionellen Seite kann dies beispielsweise dann der Fall sein, wenn sie unbesehen zu einer bloßen Übernahme algorithmisch

374 Treuthardt/Kröger 2019.
375 https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing [17.01.2023].
376 Hartmann/Wenzelburger 2021.
377 https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing [17.01.2023].

vorgeschlagener Ergebnisse führt – aus welchen Gründen auch immer (z. B. Zeitdruck, allmähliche Gewöhnung oder um im Falle einer juristischen Auseinandersetzung auf der vermeintlich sicheren Seite zu sein). Auch für die von den Ergebnissen betroffenen Personen sind negative Effekte möglich, etwa wenn ihnen aufgrund algorithmisch unterstützter Entscheidungen Handlungs- oder Entwicklungsmöglichkeiten genommen werden.

So dokumentieren die hier vorgestellten Einsatzmöglichkeiten prädiktiver Diagnosen von Gefährdungspotenzialen in den Bereichen Kindeswohlgefährdung und Bewährungshilfe auf der einen Seite einen gewissen Zugewinn an Sachangemessenheit und Wirksamkeit von sozialprofessionellen Interventionen, wenn die Entscheidung der Fachkräfte durch algorithmenbasierte, statistisch-aktuarische Eintrittswahrscheinlichkeitsberechnungen unterstützt wird. Auf der anderen Seite offenbaren sich gleichwohl auch Schwachstellen und Risiken für eine sachgemäße und nicht zuletzt personenorientierte Entscheidung, die den individuellen Hilfebedarfen der Betroffenen Rechnung zu tragen hat.

Hier treten wieder die beiden zuvor genannten Problemfelder auf: Einerseits können sich bei datenbasierten Systemen diskriminierende Effekte zeigen (Algorithmic Bias). So kann bereits die Datenlage, die den standardisierten Entscheidungen und mehr noch den digitalisierten algorithmischen Prozessen der Mustererkennung zugrunde liegt, die Sachlage verfälschen und gefährliche Stereotypisierungen und gesellschaftliche Ungleichheit aus der Vergangenheit in die Gegenwart und Zukunft verlängern. Wenn beispielsweise bei der prognostischen Einschätzung der Kindeswohlgefährdung auf Prädiktoren zurückgegriffen wird, die – wie das Beziehen von Sozialhilfe, alleinerziehende Elternschaft oder früherer Kontakt mit einer Jugendschutzbehörde – ohne Beachtung des konkreten Kontextes höchst zweifelhaft sind, dann werden sie gleichsam automatisch zukünftige Entscheidungen fehlleiten sowie bestehende Ungleichheiten (Beziehen von Sozialhilfe)

und Vorurteilsstrukturen (alleinerziehende Elternschaft, vorurteilsbelastete oder sogar bösartige Fehlanzeigen bei Kinderschutzbehörden) unbesehen fortschreiben.[378]

Andererseits steigt das Risiko eines Automation Bias mit Auswirkungen auf Fragen der Autonomie, Autorschaft und Verantwortung.[379] Gerade bei Entscheidungen, die mit einer großen prognostischen Unsicherheit konfrontiert sind und zugleich gravierende Auswirkungen haben, besteht die latente Tendenz, automatisierten Entscheidungsprozeduren mehr zu vertrauen als menschlichen Entscheidungen und damit die Verantwortung – zumindest unbewusst – von sich auf diese „Quasi-Akteure" zu delegieren. Bisweilen wird versucht, dieser Gefahr vorzubeugen, indem bei Verwendung eines Entscheidungsunterstützungstools ein entsprechender Warnhinweis gegeben wird.[380] Eine weitere denkbare Vorkehrung wäre die Verpflichtung der entscheidenden Fachkräfte, die etwaige Übernahme des algorithmischen Entscheidungsvorschlags – etwa mit Verweis auf die eigene erfahrungsbezogen intuitive oder kollegial erörterte Prognose – ausdrücklich zu begründen.

Mit Blick auf die Bewährungshilfe birgt die Nutzung statistisch-aktuarischer Risikoprofile darüber hinaus die Gefahr, dass die eigentlich primäre Zielsetzung der Bewährungshilfe, nämlich die subjektiven wie die objektiven Ressourcen der bewährungspflichtigen Person für Rehabilitation und Wiedereingliederung in die Gesellschaft zu aktivieren, zurückgedrängt wird, zugunsten der Risikominimierung im Bereich der Gefährdung für andere.[381] Diese Verschiebung geht eindeutig zulasten der bewährungspflichtigen Person und

378 Gillingham 2021, 32.
379 Safdar/Banja/Meltzer 2020.
380 In den USA wird bei Verwendung des COMPAS-Systems zur Risikoprognose von (ehemaligen) Strafgefangenen ein entsprechender Warnhinweis gegeben. Über die Wirksamkeit dieses Warnhinweises liegen bislang aber keine Erkenntnisse vor (Butz et al. 2021, 252).
381 Ghanem 2021.

vermindert im Falle der Nichtaktivierung ihrer Ressourcen ihre Lebensführungskompetenz.[382]

Insgesamt zeigt sich eine ambivalente Entwicklung für die Entscheidungshoheit der eingesetzten Fachkräfte: Algorithmenbasierte Entscheidungshilfen sollen lediglich fachlich weniger angemessene, intuitive Entscheidungsgrundlagen ersetzen und die Basis für qualitativ bessere Entscheidungen der Fachkräfte liefern, welche jedoch die letztverantwortliche Entscheidungsbefugnis beibehalten. Faktisch aber können die algorithmenbasierten Entscheidungshilfen eine präfigurative Wirkmacht entfalten, die die letztendscheidenden Fachkräfte schon in ihrer Wahrnehmung von Hilfebedarfen in vorgegebene Deutungs- und Entscheidungsmuster lenkt. Dieser Effekt wird verstärkt, wenn sich Fachkräfte für eine von den digital vorgeschlagenen Optionen abweichende Entscheidung rechtfertigen müssen – etwa vor Vorgesetzten oder gar vor Gerichten. Damit kann es im faktischen Verhalten zu einer Verminderung der Entscheidungskompetenz der Fachkräfte und mithin zu ihrer schleichenden Verdrängung oder sogar Ersetzung kommen. Die Entscheidungskompetenz der Fachkräfte würde prekär: formal und prima facie gegeben, faktisch aber unterwandert durch die Wirkmacht algorithmenbasierter Entscheidungsvorgaben.

Auch in anderen Bereichen der öffentlichen Sozialverwaltung kommt es zu ähnlichen Entwicklungen, wenn algorithmische Systeme zur Prognose des mutmaßlichen Entwicklungspotenzials von Hilfebedürftigen eingesetzt werden, beispielsweise bei Entscheidungen über die Gewährung von Sach-, Geld- und Dienstleistungen (z. B. Sozialhilfe oder Weiterbildungsmaßnahmen für Erwerbslose). Algorithmenbasierte Prognosen und Klassifizierungen können hier unmittelbare Folgen für die Art und das Ausmaß von Unterstützungsmaßnahmen haben und damit entscheidend für die Gewährung oder auch Verweigerung von Lebenschancen sein. Dies kann

382 Butz et al. 2021, 254 ff.

zur Verminderung von Entwicklungsmöglichkeiten und Autorschaft der betroffenen Personen führen.

Die professionell ausgestaltete dialogisch-interaktive Beziehungsarbeit bildet den zentralen Rahmen zur Identifikation eines individuellen Hilfebedarfs. Der Einsatz standardisierter Erfassung von Hilfebedarfen birgt demgegenüber das Risiko der Entkopplung aus einer dialogischen Beziehungsarbeit, jedenfalls dann, wenn nicht explizit dialogische Konzepte weiterhin vorgesehen bleiben. Der Hilfebedarf erschließt sich nicht allein aus objektiven Tatbestandsmerkmalen, sondern muss aus fachlichen Gründen die subjektive Perspektive der hilfebedürftigen Person auf die eigene Lebenslage einbeziehen. Eine Einbindung der betroffenen Person ist auch entscheidend für die Erfahrung von Selbstwirksamkeit[383], ohne die positive Effekte selbst materieller Hilfeleistungen schnell verpuffen und damit kaum nachhaltig wirken. Die Gefahr, dass die Individualität von Klientinnen und Klienten ignoriert und damit die Wirksamkeit der angebotenen Unterstützung verfehlt wird, steigt im Zuge algorithmenbasierter Informatisierung des Sozialwesens erheblich an.[384]

Ein eindrückliches Beispiel für die letztgenannte Problematik algorithmischer Chancenprognose ist das Arbeitsmarktchancen-Assistenzsystem (AMAS), auch AMS-Algorithmus genannt, in Österreich.[385] Es kommt in der Sozialverwaltung dann zum Einsatz, wenn bei einer arbeitssuchenden Person die Chancen ermittelt werden, innerhalb eines bestimmten Zeitfensters erfolgreich wieder in ein reguläres Erwerbsarbeitsverhältnis zu gelangen. Das System klassifiziert alle Arbeitssuchenden in Gruppen mit niedriger, mittlerer oder hoher Erfolgsprognose. Dementsprechend werden ihnen bestimmte Service-, Beratungs- oder Betreuungsleistungen angeboten oder aber auch nicht angeboten. Zwar ist die Entscheidung

383 Beck/Greving 2012; Grunwald/Thiersch 2016.
384 Ley/Reichmann 2020; Görder 2021.
385 Allhutter et al. 2020a.

der Fachkräfte formal vom Vorschlag des Systems abgetrennt. Personen, die mit ihrer Entscheidung von der algorithemenbasierten Einstufung abweichen und eine Umstufung vornehmen, müssen dies aber ausdrücklich vermerken und begründen. Der AMS-Algorithmus – so eine interdisziplinäre Studie des Instituts für Technikfolgen-Abschätzung der Österreichischen Akademie der Wissenschaften – ist ganz offensichtlich an den Werten, Normen und Zielen einer restriktiven Fiskalpolitik ausgerichtet.[386] Ein solches System ist mitnichten neutral. Vielmehr läuft es den Zielen eines personenorientierten Hilfesystems, das individuelle Hilfebedarfe betroffener Personen fokussieren muss, diametral zuwider.[387] Mit dem Einsatz solcher Systeme ist oftmals die Abkehr von einem zentralen Prinzip des Sozial- und Wohlfahrtswesens verbunden: der Ermittlung des individuellen Hilfebedarfs und der auf diesem Hilfebedarf aufruhenden Angebote, neben sächlicher und finanzieller Unterstützung besonders durch Dienstleistungen wie Weiterbildungsangebote jeder hilfebedürftigen bzw. leistungsberechtigten Person eine Lebensführung zu ermöglichen, die „der Würde des Menschen entspricht" (§ 1 Abs. 1 SGB II).

Angesichts der zuvor geschilderten Problematik ist es nicht verwunderlich, dass die Automatisierung der öffentlichen Verwaltung im Sozialwesen häufig kritisch gesehen wird. So wird in einem vielbeachteten Sonderbericht der Vereinten Nationen vor der Gefahr einer „digitalen Wohlfahrtsdystopie", das heißt vor der Schaffung einer menschenrechtsfreien Zone gewarnt, die insbesondere dann problematisch sei, wenn private Unternehmen für die Entwicklung und Implementierung von IT-Lösungen für den Sozialstaat verantwortlich zeichnen.[388] Diese Kritik ist keine Einzelmeinung. In den letzten Jahren haben sich immer mehr journalistische, wissenschaftliche und zivilgesellschaftliche Arbeiten und Initiativen kritisch mit

386 Ebd.
387 Allhutter et al. 2020b.
388 Alston 2019, 2.

den Auswirkungen staatlicher Automatisierungsbemühungen im Sozialwesen befasst. So kommt etwa Virginia Eubanks zu dem Schluss, dass automatisierte Entscheidungssysteme das soziale Sicherheitsnetz in den USA unterminieren, mit einem Vertrauensverlust dem Staat gegenüber einhergehen und somit Armut, Ungleichheit und Ungerechtigkeit weiter verstärken würden.[389] Zu einem ähnlichen Verdikt kommt Monique Mann im Zuge ihrer Analyse des „Robodebt Scheme", eines automatisierten Inkassosystems der australischen Sozialhilfe, das gefährdete und schutzbedürftige Bevölkerungsgruppen unrechtmäßig mit Schulden belaste, anstatt Armut und Ungleichheit zu verringern.[390]

Während Automatisierungsprojekte von staatlicher Seite meist als Mittel zur Kostenreduktion und erhöhter Effizienz gepriesen werden, zeigen Analysen wie die oben genannten mögliche Gefahrenpotenziale und weisen auf die Notwendigkeit einer breiten öffentlichen Debatte insbesondere über die ethischen Fragen hin. Die Verminderung menschlicher Autorschaft hat viele Facetten. Aufseiten der Menschen, die für Entscheidungen verantwortlich sind, kann sie in einer Herabstufung der Betrachtung individueller Fälle zugunsten einer statistik- und datengetriebenen Klassifikation bestehen oder auch in der schleichenden Gewöhnung an ADM-Verfahren. Dadurch kann auch die Schwelle sinken, trotz eigener anderslautender Einschätzung den Empfehlungen der ADM-Systeme zu folgen und damit die Grenze von einem *decision support* zu einem *automated decision making* zu überschreiten, in dem der Maschine die Letztentscheidungsmacht überlassen wird. Aufseiten der von den Entscheidungen jeweils Betroffenen zeigt sich die vermindernde Wirkung in der Nichtbeachtung von Aktivierungspotenzialen für eine eigenständige Lebensführung bis hin zur Verweigerung eigentlich offenstehender und nutzbarer Entwicklungschancen.

389 Eubanks 2018.
390 Mann 2020.

8.4 Predictive Policing – KI in der Kriminalitätsbekämpfung

In der Kriminalitätsbekämpfung können algorithmenbasierte Risikoanalysen nicht nur – wie im Falle der Bewährungshilfe – überführte straffällige Personen betreffen, sondern auch zur Identifikation potenziell straffälliger Personen im Rahmen präventiver Polizeiarbeit eingesetzt werden. Üblicherweise werden unter dem Begriff des Predictive Policings analytisch-technische, häufig algorithmenbasierte Anwendungen verstanden, die präventive Polizeiarbeit unterstützen und mittels Prognosen künftiger Straftaten, straffälliger Personen und Tatorte der Verhinderung von Verbrechen dienen (siehe Infokasten 12).[391]

Die entsprechenden Programme kommen mit dem Versprechen daher, große Mengen an mitunter heterogenen und verstreuten Daten schneller miteinander zu verknüpfen und so zur Verbrechensbekämpfung auswerten zu können. Verwendet werden sowohl raum- als auch personenbezogene Verfahren. Raumbezogene Methoden identifizieren durch Verknüpfung diverser Datenbestände Orte, an denen mit einer bestimmten Wahrscheinlichkeit eine Straftat begangen wird, während personenbezogene Methoden mit Täterprofilen und Opfermerkmalen arbeiten, um Personen, die mit hoher Wahrscheinlichkeit straffällig werden könnten, vorab zu erkennen. In den USA kommen sie seit 2005 zum Einsatz.

Infokasten 12: Predictive Policing

In Deutschland kommen bislang in erster Linie raumbezogene Verfahren des Predictive Policings zum Einsatz, die durch die Ausweisung prädiktiver

391 Härtel 2019, 54 f.; Rademacher 2017, 368 ff. Auch wenn Bewährungsentscheidungen nicht durch Polizeibehörden, sondern Gerichte getroffen werden, lassen sich auch dort Verfahren ähnlich dem Predictive Policing verwenden, weil diese Entscheidungen einem gefahrenabwehrrechtlichen Zweck verschrieben sind. Der Sache nach geht es um eine Gefährlichkeitsprognose in Bezug auf eine Person, die bereits eine Straftat begangen hat, anlässlich dieser konkreten Tat.

Risikogebiete mit zeitlicher Präferenz gekennzeichnet sind.[392] In der Hauptsache richten sich die Verfahren auf die Vorhersage raumzeitlicher Parameter bei Wohnungseinbruchdiebstählen.[393] Dabei kommen folgende Programme zum Einsatz: PRECOBS, SKALA, KrimPro, PreMAP und KLB-operativ.[394] PRECOBS, PreMAP und KLB-operativ beruhen auf der Near-Repeat-Methode[395], basierend auf dem Umstand, dass nach einer Ausgangstat in räumlich-zeitlicher Nähe häufig weitere Taten begangen werden[396], also dass das Begehen einer Straftat weitere Taten in unmittelbarer Nähe und innerhalb eines bestimmten Zeitraums nach sich ziehen werde[397]. Dabei werden in der Regel polizeilich ermittelte Daten zugrunde gelegt.[398] SKALA hingegen legt Kriminalitätstheorien zugrunde, die nach Auffassung des nordrhein-westfälischen Landeskriminalamtes geeignet sind, ein möglichst realistisches Bild von Wohnungseinbruchdiebstählen zu zeichnen.[399] Dazu gehören Rational-Choice-Theorien, wonach Personen, die einen Einbruch planen, vor der Tat eine Kosten-Nutzen-Abwägung vornehmen[400], dass also zum Beispiel die erwartete Höhe des Diebesgutwerts die Wahrscheinlichkeit von Wohnungseinbruchdiebstählen in einem Gebiet erhöht[401]. Auch KrimPro beruht auf einem kriminologisch komplexeren Ansatz und stützt seine Prognose auf verschiedene Kriminalitätstheorien.[402]

Personenbezogene Verfahren des Predictive Policings[403] stützen die Prognose auf Täter- bzw. Opfermerkmale[404]. Zu nennen ist beispielhaft das in

392 Povalej/Volkmann 2021, 58.
393 Egbert 2018, 244.
394 Egbert 2020, 33.
395 Knobloch 2018, 13.
396 Gluba 2017, 369.
397 Kuhlmann/Trute 2021, 105.
398 Eisele/Böhm 2020, 522.
399 Landeskriminalamt NRW 2018, 10 ff.; Eisele/Böhm 2020, 523.
400 Lüdemann/Ohlemacher 2002, 54.
401 Landeskriminalamt NRW 2018, 12.
402 Knobloch 2018, 13; Eisele/Böhm 2020, 524.
403 Auch Technologien, die im Kontext der Entscheidung über die Strafaussetzung zur Bewährung zum Einsatz kommen (z. B. die Software COMPAS), lassen sich als Verfahren des Predictive Policings einstufen, insoweit diese ebenfalls auf die individuelle Gefährlichkeit der straffälligen Person abstellen. Sofern letzterer nämlich eine negative Sozialprognose erteilt wird, steht dies einer Strafaussetzung zur Bewährung entgegen. Damit ist die Bewährungsentscheidung der Sache nach eine gefahrenabwehrrechtliche. Im deutschen Recht muss gleichwohl berücksichtigt werden, dass die Tätigkeit von Gerichten, die über die Strafaussetzung zur Bewährung entscheiden, klassischerweise von der Polizeiarbeit getrennt wird. Es ist daher weniger üblich, die Bewährungsentscheidung als polizeiliche Tätigkeit einzuordnen, was zumindest formell gegen die Einstufung als Predictive Policing spricht.
404 Povalej/Volkmann 2021, 58. Die Differenzierung zwischen raum- und personenbezogenen Verfahren des Predictive Policings ist alles andere als trennscharf. Überschneidungen ergeben sich bereits daraus, dass raumzeitliche Parameter nicht selten auf die Person des jeweiligen Gefährders schließen lassen. Gleichwohl handelt es sich um eine etablierte Differenzierung, die hier ebenfalls herangezogen wird.

Großbritannien eingesetzte Programm HART[405] sowie die sogenannte Strategic Subject List der Polizei von Chicago (USA)[406]. In Deutschland wird seit 2017 vom Bundeskriminalamt das Risikobewertungsinstrument RADAR-iTE eingesetzt. Es ordnet Personen aus dem militant-salafistischen Spektrum auf Basis einer Verhaltensanalyse als „Gefährder" oder „Relevante Personen" auf einer Risikoskala ein.[407] Seit 2017 wird in Hessen das Programm hessenDATA, basierend auf der Analysesoftware Gotham des US-amerikanischen Unternehmens Palantir Technologies, im Rahmen der Terrorismusbekämpfung auf der Grundlage von § 25a Abs. 1 Alt. 1 des Hessischen Gesetzes über die öffentliche Sicherheit und Ordnung (HSOG) eingesetzt.[408] Es nutzt Informationen aus polizeilichen Datenbanken, Verkehrsdaten aus der Telekommunikationsüberwachung und von Telekommunikationsanbietern zur Verfügung gestellte Daten. Einbezogen werden außerdem sogenannte forensische Extrakte wie zum Beispiel die Ergebnisse der Beschlagnahme eines Mobiltelefons und Informationen aus sozialen Netzwerken.[409] Eine Rechtsgrundlage für personenbezogenes Predictive Policing liefert zudem § 4 des Fluggastdatengesetzes (FlugDaG). Darin wird das Bundeskriminalamt ermächtigt, Fluggastdaten[410] automatisiert nach bestimmten Mustern abzugleichen. Dies dient der Identifikation von Personen, bei denen ein gewisses Risiko für die Begehung einer terroristischen Straftat oder einer Straftat der schweren Kriminalität besteht.[411] Nachdem das Bundesverfassungsgericht unter anderem die hessische Regelung in § 25a Abs. 1 Alt. 1 HSOG für verfassungswidrig erklärt und eine Neuregelung verlangt hat, weil sie keine ausreichende Eingriffsschwelle enthalte[412], müssen die einschlägigen Ermächtigungsgrundlagen auch in anderen Gesetzen an die vom Bundesverfassungsgericht formulierten Vorgaben angepasst werden.

Der Einsatz digitaler Technologien des Predictive Policings – vor allem im Hinblick auf personenbezogene Verfahren – wird kontrovers diskutiert. Einerseits geht damit die Hoffnung

405 Oswald et al. 2018.
406 Sommerer 2020, 80 ff.
407 https://www.bka.de/DE/Presse/Listenseite_Pressemitteilungen/2017/Presse2017/170202_Radar.html [17.01.2023].
408 Sommerer 2020, 90.
409 Hessischer Landtag 2019, 18 f.
410 Fluggastdaten i.S.v. § 2 Abs. 2 FlugDaG sind unter anderem Anschrift und Kontaktangaben der Reisenden (Nr. 5), vollständige Gepäckangaben (Nr. 7), Informationen über nicht angetretene Flüge (Nr. 14) und Angaben über Mitreisende (Nr. 19). Gemäß § 4 Abs. 3 FlugDaG werden die Muster für den Abgleich von der Fluggastdatenzentralstelle des Bundeskriminalamts unter Einbeziehung des Datenschutzbeauftragten erstellt. Allerdings ist nach wie vor ungeklärt, welche Daten in die Mustergewinnung einbezogen werden (Sommerer 2020). Als Beispiele für Muster, die über Fluggäste gewonnen werden, sind folgende Fälle zu nennen: Die Buchung des Fluges erfolgte kurzfristig; es wurde in bar gezahlt; bisher nicht alleinreisende Minderjährige reisen nunmehr allein (Münch, in: Deutscher Bundestag 2017, 26).
411 Sommerer 2020, 96.
412 Urteil vom 16. Februar 2023 – 1 BvR 1547/19, 1 BvR 2634/20 (NJW 2023, 1196).

auf eine Verbesserung der polizeilichen Arbeit und damit den besseren Schutz möglicher Opfer einher (vgl. das Beispiel zum Onlinegrooming weiter unten).[413] Andererseits ist algorithmenbasierte Verbrechensbekämpfung bzw. -verhinderung mit verschiedenen Risiken und Problemen verbunden. Diese betreffen zunächst Fragen der Genauigkeit bzw. Fehlerhaftigkeit von Vorhersagen durch KI-Systeme. Damit verknüpft ist die Frage, ob solche Vorhersagen für verschiedene Personengruppen gleichermaßen zuverlässig funktionieren. Darüber hinaus besteht die Gefahr, dass Personen aufgrund ihrer Herkunft, ihres Wohnorts oder ihrer Vorgeschichte diskriminiert werden.[414] Das tatsächliche Vorliegen einer ungerechtfertigten Ungleichbehandlung hängt maßgeblich von den verwendeten Parametern, deren Gewichtung innerhalb der Prognoseentscheidung, den verfügbaren Daten und deren Charakteristik ab.[415] Von nicht nur technischer, sondern auch ethischer Relevanz sind hierbei auch methodische Entscheidungen, die das Ausmaß und Verhältnis falsch-positiver und falsch-negativer Ergebnisse vorbestimmen. Auch wenn Fehler selbstverständlich auch bei menschlichen Urteilen auftreten, so besteht bei Entscheidungsunterstützung durch Software eine noch größere Gefahr, dass Fehler und Verzerrungen systembedingt besondere Breitenwirkung entfalten. Gerade bei polizeilichen Maßnahmen wirken beide Arten von Fehlern folgenschwer – sowohl wenn fälschlicherweise eine polizeiliche Maßnahme gegen eine Person ergriffen wird als auch wenn ein notwendiger polizeilicher Zugriff auf eine Person unterbleibt. Gesellschaftlich muss also entschieden werden, in welchem Umfang entsprechende Risiken hinnehmbar sind, und auch, ob es

413 Ob eine solche tatsächlich gelingt, ist umstritten (Eisele/Böhm 2020, 531 f.).

414 Zu diskriminierenden und stigmatisierenden Effekten vgl. Egbert/Leese 2020, 186 ff.

415 Beispielsweise lässt sich darüber streiten, ob es per se diskriminierend ist, im Rahmen einer kriminalrechtlichen Prognose den Wohnort oder den sozioökonomischen Status heranzuziehen, wenn diese Umstände zugleich empirisch belegt kriminalitätssteigernde Faktoren darstellen (vgl. ausführlich – auch mit Blick auf Predictive Policing – Kischel, in: BeckOK GG, 52. Edition, Stand: 15.08.2022, Art. 3 Rn. 218d).

einen Unterschied macht, wenn der Fehler auf der Technik oder auf menschlichem Versagen beruht.

Eine weitere Problematik, die Gegenstand intensiver Debatten ist, ist die Frage des Schutzes der Privatsphäre im Kontext von Predictive Policing im Allgemeinen und in Bezug auf Chatkontrollen im Besonderen (siehe Infokasten 13).[416] Die für die Polizeiarbeit herangezogenen Daten sind in aller Regel besonders sensibel. Bei sogenannten Chatkontrollen zur Prävention und Bekämpfung des sexuellen Missbrauchs von Kindern, zu denen die Europäische Kommission im Mai 2022 einen Verordnungsvorschlag (COM/2022/209)[417] vorgelegt hat, geht es auch um die Frage, ob eine anlasslose und flächendeckende Überwachung privater Kommunikation gerechtfertigt werden kann, gerade auch in Anbetracht vorhandener Kritik bezüglich der Effektivität solcher Kontrollen für den Schutz der Kinder und der generellen Kritik an Maßnahmen der Vorratsdatenspeicherung. Nach Einschätzung des Bundesdatenschutzbeauftragten stellt der Verordnungsentwurf einen unverhältnismäßig intensiven Eingriff in die Grundrechte dar.[418]

Infokasten 13: Onlinegrooming

In den sozialen Medien und dem Internet sind Kinder und Jugendliche dem Risiko des sogenannten Onlinegroomings ausgesetzt, bei dem Erwachsene eine emotionale Beziehung zu Minderjährigen mit dem Ziel des sexuellen Missbrauchs aufbauen.[419] Die Europäische Kommission hat im Mai 2022 einen Verordnungsentwurf zur Prävention und Bekämpfung des sexuellen Missbrauchs von Kindern vorgelegt. Dieser sieht die systematische Identifizierung entsprechender Anbahnungen in Chats als präventives Mittel vor,

416 Singelnstein 2018.
417 https://eur-lex.europa.eu/legal-content/DE/TXT/?uri=CELEX:52022PC0209 [17.01.2023].
418 https://www.bfdi.bund.de/SharedDocs/Pressemitteilungen/DE/2022/09_Chatkontrolle.html?nn=252104 [02.03.2023]; vgl. auch Entwurf einer Stellungnahme der Bundesregierung vom 17. Februar 2023 zum Vorschlag einer CSA-Verordnung der EU-Kommission, veröffentlicht auf netzpolitik.org: https://netzpolitik.org/2023/positionspapier-innenministerium-macht-wenig-zugestaendnisse-bei-chatkontrolle/#2023-02-17_BMI_Chatkontrolle_Entwurf [02.03.2023].
419 Wachs/Wolf/Pan 2012.

um potenzielle Straffällige noch vor der Tat zu entdecken und die Minderjährigen rechtzeitig vor der Sexualstraftat zu schützen.[420] Um Grooming früh und genau zu identifizieren, sollen Chats online bereits mittels automatisierter und spezifisch für diese Problematik trainierter KI-Algorithmen analysiert werden. Die von den Algorithmen zu lösende Aufgabe ist komplex, da die Missbrauchsstrategie sich erst über einen längeren Zeitraum hinweg entfaltet, vom Aufbau von Vertrauen und dem Austausch persönlicher Informationen bis hin zur verharmlosenden Kontextualisierung sexueller Handlungen und schließlich der Vereinbarung persönlicher Treffen.

Die Technologie ist stark umstritten. Aufseiten der Chancen wird insbesondere die Verhinderung von Gewalttaten und sexuellen Übergriffen an Kindern und somit der Schutz einer besonders vulnerablen Bevölkerungsgruppe in den Vordergrund gerückt. In der Kritik stehen hingegen die Risiken einer anlasslosen und flächendeckenden Überwachung privater Kommunikation und es gibt Zweifel an der Genauigkeit und Unverzerrtheit der Analysen sowie der Effektivität der Maßnahmen in Bezug auf den Schutz von Kindern.

Zu beachten ist auch der Zielkonflikt zwischen der Genauigkeit und der Schnelligkeit einer Vorhersage, dessen ethische Implikationen situationsadäquat abgewogen werden müssen. Sowohl falsch-positive als auch falsch-negative Vorhersagen, bei denen also Personen fälschlicherweise als Straffällige klassifiziert werden bzw. tatsächlich straffällige Personen nicht detektiert werden, können gravierende Auswirkungen für die fälschlich Verdächtigten oder für die fälschlich nicht geschützten Kinder haben. Eine KI-basierte Detektion kann daher zunächst nur als Alarmzeichen dienen, das die Überprüfung der Chatinhalte durch menschliche Sachverständige erzwingt. Im Falle einer überprüften Auffälligkeit muss vom Anbieter der Chatplattform sichergestellt werden, dass die Informationen auch an die Strafverfolgungsbehörden gemeldet werden.

Der Erfolg von KI-Algorithmen bei der rechtzeitigen Identifizierung Krimineller hängt auch vom Zugang zu Trainingsdatensätzen und deren Qualität ab, ebenso wie von der Effizienz der Algorithmen bei limitierter Rechenkapazität, da oftmals eine zeitnahe, lokale Analyse auf mobilen Endgeräten (wie etwa Smartphones) erforderlich ist.[421] Es ist zu erwarten, dass Algorithmen fortwährend einer Aktualisierung auf der Grundlage neuer Trainingsdaten bedürfen, wenn diejenigen, die solche Sexualstraftaten planen und begehen, Umgehungsstrategien (wie etwa adaptierte sprachliche Äußerungen) entwickeln.

Personenbezogene Verfahren des Predictive Policings können, wie in anderen Feldern auch, menschliche Beurteilung und Entscheidung unterstützen und dadurch Handlungsmöglichkeiten insbesondere der Strafverfolgungsbehörden erweitern.

420 Vogt/Leser/Akbik 2021.
421 Ebd.

Andererseits können sie die Handlungsmöglichkeiten der verschiedenen Beteiligten aber auch vermindern oder – im Falle einer vollständigen Übertragung von Entscheidungen an algorithmische Systeme – gar eliminieren. Wenn bei der Verhinderung einer Straftat hohe Anforderungen an Schnelligkeit bestehen, kann es zu einem Konflikt mit dem gleichzeitigen Gebot von möglichst großer Sorgfalt und Heranziehung menschlicher Beurteilung kommen. Nicht zuletzt wird die Sorge geäußert, dass mit algorithmengesteuerter Polizeiarbeit das Risiko der Verfestigung eines mechanischen Menschenbildes einhergehen könne, das den einzelnen Menschen verobjektiviere, seine Individualität auf eine datengetriebene Klassifikation reduziere, jedoch die gesamtgesellschaftlichen Ursachen von Kriminalität unberücksichtigt lasse.[422]

Der Einsatz von algorithmischen Verfahren des Predictive Policings kann die Versuchung bzw. Erwartung einer möglichst weitgehenden, gar lückenlosen Vermeidung von Straftaten im Vorfeld ihrer Begehung wecken. Für eine weiter voranschreitende Durchdringung des Lebensalltags der Menschen mit gefahrenabwehrrechtlichen Maßnahmen sind im Zeitalter der Digitalisierung die technischen Voraussetzungen erfüllt. Immer wieder wird dann auch befürchtet, die freiheitliche Ordnung würde allmählich, mit dem Argument von mehr Sicherheit im Rücken, einem Überwachungsstaat weichen, in dem lückenlose staatliche Überwachung wenig bis keinen Raum für individuelle Freiheit lassen würde. Ist diese Sorge zwar grundsätzlich verständlich, weil der Preis für die erzielte Sicherheit und Gefahrenabwehr zu hoch wäre[423], darf sie jedoch nicht dazu führen, Pauschalurteile gegen Optionen wie das Predictive Policing zu fällen. Wie in diesem Kapitel an vielen Beispielen gezeigt wurde, hängt es vom Design der technischen Systeme und vor allem von ihrer organisatorischen Einbettung in menschliche Entscheidungsverfahren ab, ob

422 Zu weiteren Bedenken siehe Egbert 2018, 256.
423 Rostalski 2019, 484 f.

die Autorschaft der verschiedenen Beteiligten oder Betroffenen erweitert oder vermindert wird. Statt eines Pauschalurteils bedarf es also der am Ideal menschlicher Urteilsfähigkeit und Verantwortung ausgerichteten Gestaltung der Systeme und ihres Einsatzes in der polizeilichen Praxis.

Dies entspricht dem Bild menschlicher Autorschaft, die prinzipiell dazu in der Lage ist, sich auf der Basis von Einsicht in die Richtigkeit eines bestimmten Verhaltens, beispielsweise zur Wahrung des Körperverletzungsverbots, selbstständig und eigenverantwortlich für die Befolgung des Rechts zu entscheiden. Würde diese Entscheidungsfreiheit in der Folge umfassender staatlicher Sicherungsinstrumente abhandenkommen, würden sich Einzelne nicht länger als verantwortlich für das Funktionieren des freiheitlichen Miteinanders begreifen – mit all den Negativfolgen, die es für ein Gemeinwesen hat, wenn sich dessen Mitglieder selbst nicht mehr in der Pflicht sehen, in der Gestaltung des Rechts und seiner Wahrung eine aktive Rolle einzunehmen.

8.5 Fazit und Empfehlungen

Die Digitalisierung der öffentlichen Verwaltung wird seit über zwanzig Jahren zumeist unter Aspekten von Serviceorientierung, Modernität sowie Effizienz und Effektivität thematisiert. KI und dadurch ermöglichte automatisierte Entscheidungsverfahren führen zu neuen Möglichkeiten und Herausforderungen, die erheblich weitreichende ethische und demokratietheoretische Fragen aufwerfen. Denn da staatliches Handeln von vielen Menschen zu einem großen Teil über die öffentliche Verwaltung erfahren wird, beispielsweise in Finanzverwaltung, Meldebehörden oder Sozialwesen, ist mit der Einführung beispielsweise automatisierter Entscheidungsverfahren unmittelbar das Verhältnis von Bürgerschaft und Staat betroffen. Dies gilt etwa in Bezug auf Nachvollziehbarkeit, Erklärbarkeit und Vertrauenswürdigkeit im Verwaltungshandeln, aber auch

in Bezug auf Sorgen um Diskriminierung und Technokratie, in der menschliche Kommunikation und Abwägung hinter anonymen Datenmengen und standardisierten Bedienoberflächen verschwinden. Eine in diesen Hinsichten als legitim anerkannte Verwaltung ist für ein funktionierendes Gemeinwesen und die Akzeptanz von Demokratie und Staat wesentlich. Die Erfüllung dieser auch für den Einsatz von KI in der öffentlichen Verwaltung geltenden Anforderung hängt mit den durch Automatisierung veränderten Möglichkeiten menschlicher Autorschaft im Verwaltungshandeln zusammen und führt zu den ethischen Fragen nach ihrer Erweiterung oder Verminderung durch die Einführung KI-gestützter Systeme in diesem Feld.

Seit einigen Jahren kommen in vielen Staaten Systeme zur Entscheidungsunterstützung in der öffentlichen Verwaltung zusehends zum Einsatz, so etwa zur Bewertung von Arbeitsmarktchancen, zur Prüfung und Vergabe von Sozialleistungen und bei Sicherheitsorganen. Aufgrund des dadurch ermöglichten Rückgriffs auf große Datenmengen und ihrer zielgerichteten Auswertung in kurzer Zeit können für die erforderlichen Entscheidungen bessere Grundlagen geschaffen und kann die menschliche Autorschaft unterstützt werden. Insofern also die letztliche Entscheidung, etwa über die Bemessung von Sozialleistungen oder die Rückfallwahrscheinlichkeit einer straffälligen Person, bei den Personen, die für diese Entscheidung zuständig sind, verbleibt, ist der Einsatz der ADM-Systeme in ethischer Hinsicht grundsätzlich zu begrüßen. Entscheidend ist hier, dass die Zwecke ihrer Einführung in Übereinstimmung mit dem Ziel der Erweiterung menschlicher Autorschaft stehen und die Systeme nicht etwa bloßer Effizienzsteigerung des behördlichen Funktionierens oder der Einsparung von Personal dienen.

Jedoch kann es durch Gewöhnung, Überlastung und Entscheidungsdruck zu einer allmählichen Verschiebung kommen, in deren Verlauf die Ergebnisse der ADM-Systeme von den Menschen, die eine Entscheidung treffen, zunehmend

ohne weitere Prüfung und Reflexion übernommen würden (Automation Bias). Auf diesem Weg würde menschliche Autorschaft allmählich und fast unmerklich verschwinden. Zurückbleiben würde ein automatisiertes Geschehen, in dem technische Systeme für Betroffene weitreichende, teils existenzielle Festlegungen treffen. Dies ist nicht bloß eine abstrakte Befürchtung, sondern spiegelt Erfahrungen im alltäglichen Verwaltungshandeln wider. Es kommt also zu Konflikten zwischen Erwartungen an die Erweiterung menschlicher Autorschaft einerseits und der realen Umsetzung, in der ihre Verminderung oder gar Ersetzung nicht fernliegen, andererseits. Ein empirisches Monitoring der realen Adaptation von ADM-Systemen und ihren Folgen ist notwendig, um Fehlentwicklungen in ethischer Hinsicht frühzeitig erkennen und Gegenmaßnahmen einleiten zu können.

Ferner ist an die Forderung nach Diskriminierungsfreiheit des öffentlichen Verwaltungshandelns und seiner möglichst weitgehenden Umsetzung zu erinnern. Automatisiertes Entscheiden, etwa im Sozialwesen oder im Predictive Policing, kann durch die verwendeten Datensätze und zur Auswertung herangezogenen Algorithmen diskriminierend für bestimmte Bevölkerungsgruppen wirken. Dieses mittlerweile gut bekannte Risiko[424] stellt kein Pauschalargument gegen ADM-Systeme in der öffentlichen Verwaltung dar, zumal auch menschliche Entscheidungen vor Diskriminierung nicht gefeit sind, sondern hat zur Folge, dass sowohl das Design der Systeme als auch die Trainingsdaten und dann auch ihre realen Operationen daraufhin kritisch beobachtet werden müssen.

Schließlich ist festzuhalten, dass die von Ergebnissen der ADM-Systeme Betroffenen, etwa bei Entscheidungen des Jugendamtes oder in der Steuerbemessung, die gleichen Rechte etwa auf Erläuterung und Widerspruch haben wie bei Institutionen ohne diese Systeme.

424 Orwat 2019.

Pauschale Aussagen für oder gegen KI bzw. ADM-Systeme in der öffentlichen Verwaltung sind also nicht sinnvoll. Es muss kontextbezogen im Detail eingeschätzt und abgewogen werden, welche Auswirkungen eine entsprechende Maßnahme auf die Autorschaft unterschiedlichster Beteiligter und Betroffener hat, welche Konflikte auftreten und wie mit ihnen umgegangen werden kann oder soll.

Empfehlungen

- *Empfehlung Verwaltung 1:* Die mit automatisierten Entscheidungshilfen (ADM-Systeme) einhergehende verstärkte Standardisierung und pauschale Kategorisierung von Einzelfällen müssen umso stärker hinterfragt und um spezifisch einzelfallbezogene Erwägungen ergänzt werden, je intensiver die betroffene Entscheidung individuelle Rechtspositionen berührt.
- *Empfehlung Verwaltung 2:* Es müssen geeignete technische und organisatorische Instrumente zur Vorkehrung gegen die manifeste Gefahr eines Automation Bias bereitgestellt werden, die es den Fachkräften erschweren, selbst bei einer Letztentscheidungskompetenz der algorithmischen Entscheidungsempfehlung unbesehen zu folgen. Es ist zu prüfen, ob eine Umkehrung der Begründungspflicht (nicht eine Abweichung, sondern ein Befolgen ist zu rechtfertigen) hier eine geeignete Vorkehrung sein kann.
- *Empfehlung Verwaltung 3:* Aufgrund ihrer Grundrechtsbindung sind an staatliche Einrichtungen bei der Entwicklung und Nutzung algorithmischer Systeme hohe Anforderungen in Bezug auf Transparenz und Nachvollziehbarkeit zu stellen, um den Schutz vor Diskriminierung zu gewährleisten sowie Begründungspflichten erfüllen zu können.
- *Empfehlung Verwaltung 4:* Für Softwaresysteme in der öffentlichen Verwaltung müssen Qualitätskriterien verbindlich und transparent festgelegt werden (z. B. in Bezug auf Genauigkeit, Fehlervermeidung und Unverzerrtheit). Ebenso bedarf es einer Dokumentation der jeweils

eingesetzten Methoden. Diesbezüglich sollten auch aktuelle Beschaffungspraktiken, in deren Verlauf staatliche Behörden Softwarelösungen kaufen, einer kritischen Prüfung unterzogen werden.

» *Empfehlung Verwaltung 5:* Überall dort, wo algorithmische Systeme Einsatz in der öffentlichen Verwaltung finden, gilt es, Sorge zu tragen, dass die Personen, die diese Systeme anwenden, über die erforderlichen Kompetenzen im Umgang damit verfügen. Dazu gehört neben der Kenntnis der Verwendungsweisen auch das Wissen um die Limitationen und möglichen Verzerrungen, um Systeme angemessen einsetzen zu können.

» *Empfehlung Verwaltung 6:* Die Einsichts- und Einspruchsrechte Betroffener müssen auch beim Einsatz algorithmischer Systeme effektiv gewährleistet werden. Dazu bedarf es gegebenenfalls weiterer wirksamer Verfahren und Institutionen.

» *Empfehlung Verwaltung 7:* In Öffentlichkeit, Politik und Verwaltung sollte eine Sensibilisierung gegenüber möglichen Gefahren von Automatisierungssystemen, wie etwa Verletzungen der Privatsphäre oder Formen systematisierter Diskriminierung, erfolgen. Dazu gehört eine öffentliche Debatte darüber, ob es in bestimmten Kontexten überhaupt einer technischen Lösung bedarf.

» *Empfehlung Verwaltung 8:* Im Bereich des Sozialwesens ist sicherzustellen, dass ADM-Systeme elementare fachliche Standards von sozialprofessionellen Interaktionen (z. B. gemeinsame Sozialdiagnose oder Hilfeplanung als *Teil* therapeutischer bzw. unterstützender Hilfeleistung) nicht unterlaufen oder verdrängen. Dies beinhaltet insbesondere Maßnahmen, die Vergröberungen individueller Fallkonstellationen und -prognosen durch die ADM-induzierte grobklassifizierende Einteilung von Fall- und/oder Leistungsberechtigtengruppen verhindern. Dabei ist Sorge zu tragen, dass die Feststellung individueller Hilfebedarfe nicht erschwert wird und es zu keiner schleichenden

Aushöhlung der sozialrechtlich gebotenen Identifizierung individueller Hilfebedarfe zugunsten einseitiger externer Interessen an Gefahrenminimierung oder Kostenersparnis kommt.

» *Empfehlung Verwaltung 9:* Die Arbeit von Gefahrenabwehrbehörden einschließlich der Polizei betrifft besonders grundrechtssensible Bereiche. Dies wirkt sich auf die Reichweite eines zulässigen Einsatzes von algorithmischen Systemen in der prädiktiven Polizeiarbeit aus. Risiken wie Verletzungen der Privatsphäre oder potenziell unzulässige Diskriminierungen der von dem Einsatz betroffenen Personen müssen mit Chancen auf erhebliche Verbesserungen der staatlichen Gefahrenabwehr sorgfältig abgewogen und in ein angemessenes Verhältnis gebracht werden. Hierfür erforderliche gesellschaftliche Aushandlungsprozesse sollten umfangreich geführt werden. Dabei ist der diffizilen Bestimmung des Verhältnisses von Freiheit und Sicherheit Rechnung zu tragen. Jegliche Gesetzesübertretung zu verhindern, wäre mit rechtsstaatlichen Mitteln nicht möglich.

TEIL III: QUERSCHNITTSTHEMEN UND ÜBERGREIFENDE EMPFEHLUNGEN

9 ZUSAMMENFASSUNG DER BISHERIGEN ANALYSE

9.1 Anthropologische und ethische Grundorientierung

Der Begriff der Künstlichen Intelligenz hat in der öffentlichen Debatte zunehmend an Aufmerksamkeit gewonnen und wird teils mit überzogenen Hoffnungen, teils aber auch mit fehlgeleiteten Befürchtungen verknüpft. Insbesondere in den Debatten rund um die sogenannte starke KI wird zuweilen sogar nahegelegt, maschinelle Akteure könnten normativ menschlichen Akteuren ähnlich oder gleichgestellt sein – was etwa zu Diskussionen um „Menschenrechte für Roboter" führt.

Der Deutsche Ethikrat geht in seiner Stellungnahme hingegen von einem normativ grundlegenden Unterschied zwischen Mensch und Maschine aus.[425] Softwaresysteme, auch solche, die der KI zugerechnet werden, verfügen weder über theoretische noch über praktische Vernunft. Sie können keine Verantwortung für ihr Handeln übernehmen, sie sind kein personales Gegenüber, auch dann nicht, wenn sie Anteilnahme, Kooperationsbereitschaft oder Einsichtsfähigkeit simulieren.

Anthropomorphe Missverständnisse sind möglicherweise Folge der Rede von „künstlicher Intelligenz". Der Begriff „Intelligenz" wird bislang ganz überwiegend für die Beschreibung menschlicher kognitiver Leistungen verwendet. Der Begriff der menschlichen Intelligenz verweist auf ein Ensemble von Leistungen, die eine Reihe von Primärfaktoren

425 Diese Feststellung lässt zwei Deutungen zu: Sie kann einmal im Sinne eines kategorialen Unterschieds verstanden werden, der auch durch weiteren technischen Fortschritt nicht überwunden werden kann; dieser Lesart schließt sich die Mehrheit des Deutschen Ethikrates an. Sie kann aber auch lediglich im Sinne einer Feststellung hinsichtlich des gegenwärtigen und allenfalls für die nahe Zukunft überblickbaren technischen Entwicklungsstandes verstanden werden; dieser Lesart schließt sich eine Minderheit des Deutschen Ethikrates an.

wie induktives Schließen, räumliches Vorstellungsvermögen, Wahrnehmungsgeschwindigkeit, Rechenfähigkeit, verbales Verständnis, assoziatives Gedächtnis und Wortflüssigkeit umfassen und die sich, jedenfalls auf dem heutigen Stand der Softwareentwicklung, nicht alle auf KI oder zumindest nicht vollständig auf diese übertragen lassen. Dies gilt erst recht für Konzepte wie die der sozialen oder emotionalen Intelligenz. Insofern enthält der Ausdruck „künstliche Intelligenz" eine metaphorische Verwendung von „Intelligenz", die das Missverständnis einer wenigstens weitgehenden Ähnlichkeit oder sogar Gleichheit von menschlicher Intelligenz und KI fördert. Entsprechend wäre es in vielen Fällen hilfreicher, den Begriff der Künstlichen Intelligenz nicht inflationär zu verwenden, sondern genauer zu spezifizieren, was eine Software ist, beispielsweise eine Software zur Entscheidungsunterstützung, oder was sie tut, beispielsweise basierend auf statistischen Analysen diverser Daten Prognosen erstellen.

Noch deutlicher wird ein kategorialer Unterschied zwischen Mensch und Maschine, wenn man zum Begriff der Vernunft übergeht. Die *theoretische Vernunft* richtet sich auf den Erkenntnisgewinn, um zu wahren empirischen oder aprioriischen Urteilen zu gelangen, während die *praktische Vernunft* auf ein kohärentes, verantwortliches Handeln abzielt, um ein gutes Leben zu ermöglichen. Dabei ist menschliche Vernunft als verleiblicht zu begreifen. Das heißt, die Vernunft des Menschen ist nicht hinreichend beschrieben, wenn sie als abstrakte allgemeine Menschenvernunft aufgefasst wird. Sie ist immer zugleich eingebunden in die konkrete soziale Mit- und Umwelt. Nur so ist zu erklären, dass sie handlungswirksam wird. Als solche bestimmt sie die Sozialität und Kulturalität des Menschen. „Vernünftig" handelt der einzelne Mensch als Teil einer sozialen Mitwelt und einer kulturellen Umgebung. Schon deshalb kann den in dieser Stellungnahme besprochenen Softwaresystemen weder theoretische noch praktische Vernunft zugeschrieben werden (siehe ausführliche Darstellung und Begründung in Kapitel 3).

Menschen entwickeln digitale Technik und nutzen sie als Mittel zu menschlichen Zwecken. Jedoch wirken diese Technologien zurück auf menschliche Handlungsmöglichkeiten: Dadurch können sich einerseits neue Optionen eröffnen und im günstigen Fall Freiheitsgrade vergrößern. Andererseits kann aber auch die Handlungsfähigkeit von Menschen eingeschränkt und können Anpassungen erforderlich werden, die nicht wünschenswert sind. Bei der Verwendung hoch entwickelter vernetzter Softwaresysteme haben Menschen teilweise eine Subjekt-, teilweise aber auch eine Objektrolle inne. Es kann dabei bewusst auf den Subjektstatus verzichtet werden, um Handlungsmöglichkeiten zu erweitern oder Routineaufgaben an Maschinen zu delegieren. Diese instrumentelle Funktion ist jedoch von der Selbstzwecklichkeit menschlicher Akteure zu unterscheiden. Die Fähigkeit, Urheber eigener Handlungen zu sein, die die Grundlage autonomer personaler Praxis ist, bleibt Menschen vorbehalten.

Menschen wirken zweckgerichtet und kontrolliert auf sich selbst und ihre Umwelt ein, um damit bestimmte Veränderungen zu verursachen. Dabei spielt die Möglichkeit, zwischen unterschiedlichen Handlungsoptionen wählen zu können, eine wesentliche Rolle. Auch hoch entwickelte Softwaresysteme oder KI-gestützte Roboter sind nicht in der Lage, in diesem anspruchsvollen Sinne zu „handeln". Algorithmische Systeme, die etwa bei der Auswahl von geeigneten Personen für eine Stelle eingesetzt werden, handeln oder entscheiden nicht selbst und können keine Verantwortung übernehmen. Allerdings beeinflussen sie, mitunter in hohem Maße, die Bedingungen, Möglichkeiten und Grenzen für menschliches Handeln und menschliche Verantwortungsübernahme. In den vorangegangenen Kapiteln wurde dies für die Anwendungsbereiche der Medizin, der schulischen Bildung, der öffentlichen Kommunikation und Meinungsbildung sowie der öffentlichen Verwaltung exemplarisch aufgefächert. Entscheidungsunterstützungssysteme beeinflussen menschliche Handlungsmöglichkeiten in Medizin und Verwaltung ebenso wie KI-gestützte

Lernsoftwaresysteme dies im Bildungsbereich tun oder Suchmaschinen und Empfehlungssoftware im Bereich der öffentlichen Meinungsbildung. Auch wenn Maschinen also nicht selbst handeln, so verändern sie die Handlungsfähigkeit von Menschen tiefgreifend und können Handlungsmöglichkeiten erheblich erweitern oder vermindern (vgl. Kapitel 4).

Ziel der Delegation von Handlungssegmenten an Maschinen sollte prinzipiell die *Erweiterung* menschlicher Handlungsfähigkeit und Autorschaft sein. Umgekehrt gilt es, die *Verminderung* menschlicher Handlungsfähigkeit und Autorschaft sowie eine *Diffusion* oder *Evasion von Verantwortung* zu verhindern. Entsprechend ist es erforderlich, dass die Übertragung menschlicher Tätigkeiten auf KI-Systeme gegenüber den Betroffenen transparent erfolgt. Es darf nicht zu einer Verantwortungsdiffusion kommen, die dazu führt, dass niemand mehr für Fehlentscheidungen verantwortlich ist – etwa, weil wichtige Entscheidungselemente, -parameter oder -bedingungen nicht mehr nachvollziehbar sind.

Eine zentrale Erkenntnis der hier vorgelegten Analysen lautet, dass Ausmaß und Art sinnvoller und guter Delegation nicht allgemein festgelegt werden können, sondern jeweils kontextspezifisch bestimmt werden müssen. So kann die Verminderung menschlicher Handlungsfähigkeit in bestimmten Kontexten durchaus Vorteile mit sich bringen, wenn beispielsweise durch den Verlass auf teilautomatisierte Diagnostikverfahren in der Medizin Fehldiagnosen reduziert werden. Um über Wert und Nutzen der Delegation vormals menschlichen Handelns an Maschinen zu befinden, müssen allerdings auch die langfristigen Auswirkungen dieser Delegation im Sinne einer Erweiterung oder Verminderung menschlicher Handlungsfähigkeit berücksichtigt werden. So kann sich der Verlust menschlicher Expertise, der entsteht, wenn sich bei ärztlichen Fachkräften durch den Einsatz von algorithmischen Systemen die eigene Fähigkeit zur Diagnose verringert, langfristig als Nachteil erweisen, zum Beispiel in der Ausbildung oder in

Situationen, wo die entsprechende Technik nicht (mehr) zur Verfügung steht.

Hier wird die soziale Dimension des Verhältnisses von Delegieren, Erweitern und Vermindern deutlich: Menschen können durch Prozesse des Delegierens bzw. Ersetzens unterschiedlich betroffen sein. Man denke hier etwa an den Einsatz von Entscheidungsunterstützungssystemen in der Verwaltung, der Handlungsmöglichkeiten von Fachkräften und unterschiedlichen Betroffenen auf sehr verschiedene Weise erweitern oder vermindern kann. Auch im schulischen Kontext betrifft das Delegieren von Handlungen an Maschinen Lernende anders als Lehrkräfte oder gegebenenfalls noch weitere Personen. Beide Kontexte sind von asymmetrischen Beziehungs- und Machtverhältnissen gekennzeichnet, denen bei der Bewertung der Effekte des Einsatzes von Technologien Rechnung getragen werden muss. Dies gilt umso mehr, wenn bestimmte Personengruppen aufgrund ihrer Situierung besonders vulnerabel sind, beispielsweise Kinder im Schulkontext oder gar im Falle einer Wohlgefährdung. Bei der Beurteilung technologischer Entwicklungen gilt es also, den Blick nicht nur auf diejenigen, die Technik verwenden, und direkt davon Betroffene, sondern auch auf indirekt Betroffene zu richten. Darüber hinaus gilt es, die Rolle von weiteren Beteiligten zu beleuchten, die eher im Hintergrund agieren und entweder in besonderer Weise von Einführung und Nutzung der Technologien profitieren oder durch ihre Entscheidungen zur Gestaltung oder zum Einsatz von Technik Möglichkeiten und Risiken für andere Akteure präkonfigurieren. Man denke hier sowohl an die großen Plattformen und Softwareanbieter, aber auch an Data Broker, die im Hintergrund Daten zu verschiedensten, möglicherweise auch sachfremden Zwecken verarbeiten oder verkaufen.

Die Herausforderungen stecken also wie so oft im Detail, genauer: in den Details der Technik, der Einsatzkontexte sowie der institutionellen und soziotechnischen Umgebung. Neben den allgemeinen Forderungen nach Transparenz und

Nachvollziehbarkeit, dem Schutz der Privatsphäre und der Minimierung von Diskriminierung und systematischen Verzerrungen (Bias) müssen das ethische Design und der ethische Einsatz von Technologien also immer auch die konkreten institutionellen und organisationalen Rahmenbedingungen sowie die individuellen Erfordernisse unterschiedlicher Akteure in den Blick nehmen. Im Folgenden werden die Erkenntnisse aus den Kapiteln 5 bis 8 zusammengefasst und in Kapitel 10 dann übergreifende Querschnittsthemen und Empfehlungen extrahiert. Die detaillierteren, sektorspezifischen Empfehlungen befinden sich jeweils am Ende der Kapitel 5 bis 8.

9.2 Einsichten aus den Anwendungsfeldern

Die Verwendung von Technologien verändert und erweitert die Handlungsmöglichkeiten von Menschen seit Jahrtausenden. Der Einsatz von Technologien hat dabei nicht nur intendierte Folgen im Nahbereich, sondern wirkt im Positiven wie im Negativen auch in großer räumlicher und zeitlicher Distanz. Ähnlich verhält es sich mit dem Einsatz von KI. Die Delegation von Entscheidungen, Handlungen und Handlungssegmenten an Software im Allgemeinen und KI im Besonderen schafft neue Möglichkeiten, birgt aber auch zahlreiche Risiken.

In dieser Stellungnahme hat sich der Deutsche Ethikrat exemplarisch mit einigen Anwendungen in der Medizin, der schulischen Bildung, der öffentlichen Kommunikation und der öffentlichen Verwaltung beschäftigt. Dieser Auswahl liegt die Überzeugung zugrunde, dass erst ein Blick in die jeweiligen Anwendungskontexte die besonderen Chancen und Risiken des Einsatzes von KI zutage fördert. Es wurden bewusst Sektoren ausgewählt, in denen die Durchdringung durch KI-basierte Technologien sehr unterschiedlich ausfällt. Am einen Ende des Spektrums befindet sich der Bereich der sozialen Medien mit seinen Auswirkungen auf die öffentliche Kommunikation

und Meinungsbildung. Hier sind alle, die diese Angebote nutzen, bereits jetzt in ihrem Alltag auf vielfältige Weise bei ihrer Informationsauswahl von den Auswirkungen der dort verwendeten Technologien betroffen. Am anderen Ende befindet sich der Bereich der schulischen Bildung, in dem der Einsatz von KI gegenwärtig, jedenfalls in Deutschland, noch eher die Ausnahme darstellen dürfte.

Es wurden in allen Sektoren Beispiele gewählt, die unterschiedliche Ausmaße des Ersetzens vormals menschlicher Handlungen durch KI veranschaulichen. In allen vier Sektoren gibt es Einsatzszenarien, die durch teils erhebliche Beziehungs- und Machtasymmetrien gekennzeichnet sind, was einen verantwortungsvollen Einsatz von KI und die Berücksichtigung der Interessen und des Wohls insbesondere vulnerabler Personengruppen umso wichtiger macht. Diese Unterschiedlichkeit der Art und Weise des KI-Einsatzes sowie des Ausmaßes der Delegation an Maschinen in den Blick zu nehmen, erlaubt es, nuancierte ethische Betrachtungen anzustellen.

Im besonders sensiblen Bereich der Medizin (vgl. Kapitel 5) kommen digitale Produkte mit KI-Komponenten in rasch wachsendem Umfang zum Einsatz. Dabei sind unterschiedliche Akteursgruppen zu unterscheiden, die vom KI-Einsatz unterschiedlich betroffen sind oder verschiedene Verantwortlichkeiten besitzen: diejenigen, die KI-Instrumente entwickeln, klinisch tätige Personen, Patientinnen und Patienten. Das Spektrum der Delegation an KI-Systeme reicht hier von der punktuellen Unterstützung ärztlichen Handelns über dessen ständige Begleitung bis hin zur weitgehenden oder gar vollständigen Ersetzung ärztlicher Fachkräfte durch ein KI-System. Um einen verantwortlichen Einsatz von KI-Systemen zu gewährleisten, ist es daher erforderlich, die gesamte Praxis von der Entwicklung der KI-Systeme über ihren Einsatz in der Forschung bis zur Implementierung in der medizinischen Versorgung an ethischen Standards auszurichten.

Ein wichtiges Einsatzfeld für KI-Systeme ist die Unterstützung ärztlicher Entscheidungen in der Diagnostik, da diese

große Datenmengen und eine Vielzahl relevanter Parameter auswerten und in Gestalt präziser Mustererkennung großes diagnostisches und therapeutisches Potenzial entwickeln können. Der verbesserten Diagnostik stehen allerdings auch Risiken gegenüber, etwa der Verlust eigener Diagnosekompetenz aufseiten des ärztlichen Fachpersonals im Vertrauen auf technische Systeme und die Vernachlässigung eigener Urteilskraft. Auch der Aufklärungsbedarf aufseiten der Patientinnen und Patienten steigt durch den Einsatz von KI-Instrumenten und muss berücksichtigt werden, um das Vertrauen und die personale Zuwendung im Arzt-Patienten-Verhältnis nicht zu beschädigen. Besonders ambivalent ist etwa die Ersetzung psychotherapeutischer Diagnose oder Therapie durch Chatbots. Einerseits erhofft man sich davon einen Beitrag zur Entlastung des Gesundheitssektors und einen erleichterten Zugang zur psychotherapeutischen Erstversorgung. Andererseits stellen sich zahlreiche Probleme, nicht zuletzt die Gefahr einer Personalisierung von Maschinen, die falsche Projektionen begünstigen kann.

Auch im Bereich der schulischen Bildung (vgl. Kapitel 6) kommen KI-gestützte Technologien zum Einsatz, insbesondere um personalisiertes Lernen und Lehren zu unterstützen oder Informationen zum Lerngeschehen bereitzustellen. Die maschinelle Ersetzung bestimmter pädagogischer Handlungssegmente kann auch hier sowohl zur Erweiterung, aber auch zur Verminderung der Handlungsoptionen von Lernenden und Lehrkräften führen. Die Vision einer vollständigen Ersetzung des Lehrpersonals durch Maschinen ist jedoch mit dem interpersonalen Charakter des pädagogischen Geschehens und dem Fokus auf Persönlichkeitsbildung nicht in Einklang zu bringen. Auch dürfen durch den Einsatz digitaler Tools weder die ohnehin oft dominierenden Vorstellungen einer technisch verstandenen Optimierung von Lernprozessen verstärkt noch der Blick auf das (Ab-)Prüfbare verengt werden. Im Mittelpunkt der Bildungspraxis muss vielmehr weiterhin die Herausbildung von mündigen und freien Personen stehen, die

urteilsfähig und verantwortlich handeln können. Der Einsatz digitaler Systeme muss diesen Zielen verpflichtet sein und darf nicht lediglich Selbstzweck sein. Er kann dabei je nach Ziel, Gestaltung und technischen Grundlagen sehr unterschiedlich ausfallen. So kann er etwa darauf gerichtet sein, Wissen zu vermitteln oder Fähigkeiten zu trainieren, Feedback zum Lernfortschritt zu geben oder Impulse für das pädagogische Geschehen zu geben. Dabei sollte verstärkt auf das Inklusionsziel allgemeiner Bildung geachtet werden: KI-Systeme können Schülerinnen und Schüler mit besonderen Bedürfnissen und deren individuellen Lerngeschichten adaptiv begleiten, sie unterstützen, ihre spezifischen pädagogischen Bedürfnisse zu artikulieren, oder neue Möglichkeiten eröffnen, im Krankheitsfall aus der Ferne immersiv am Schulalltag teilzunehmen. Das Unterrichtsgeschehen als Praxis der Verständigung, des Austausches von Gründen, der Stärkung der Urteilskraft und der Persönlichkeitsentwicklung sollte durch den Einsatz digitaler Tools erweitert und nicht vermindert werden. Gleichwohl können den Lehrkräften wie den Lernenden dadurch neue Handlungsoptionen eröffnet und der personalisierte Austausch vertieft werden.

Prozesse der öffentlichen Kommunikation und Meinungsbildung (vgl. Kapitel 7) sind in zunehmendem Maße durch Software im Allgemeinen und KI im Speziellen geprägt. Auf der einen Seite erweitern soziale Medien mit ihren vielfältigen und wirkmächtigen soziotechnischen Prozessen der Generierung, Kuratierung und Sortierung von Inhalten menschliche Handlungsmöglichkeiten, etwa durch den Zugang zu einer zuvor ungekannten Menge und Diversität verfügbarer Informationen und Kommunikationsmöglichkeiten. Demgegenüber steht jedoch auch die – häufig weniger sichtbare – Gefahr der Verminderung menschlicher Handlungsmacht. Suchmaschinen, Newsfeeds und Empfehlungssoftware steuern beabsichtigt oder unbeabsichtigt menschliches Verhalten, beeinflussen unseren Zugang zu Information und Kommunikation und somit zur Realität. Solche möglichen Einbußen informationeller

Selbstbestimmung sind das Resultat der zuvor geschilderten soziotechnischen und häufig opaken Prozesse. Diese Prozesse sind primär auch nicht an den Interessen der Personen, die sie zur Information und Kommunikation nutzen, ausgerichtet, sondern an denen der Plattformen selbst sowie der dort aktiven Werbetreibenden.

Von den hier behandelten Sektoren ist der Bereich von Information und Kommunikation derjenige, der am stärksten von KI durchdrungen ist und in dem sich Möglichkeiten und Risiken bereits entsprechend deutlich abzeichnen. Viele Anwendungen wären ohne den Einsatz algorithmischer Systeme, die Informationen sortieren, nicht möglich. Diese starke Durchdringung unserer Lebenswelt mit digitalen Technologien führt jedoch zu einer individuellen und kollektiven Abhängigkeit und schafft Sachzwänge, welche die Vulnerabilität der Gesellschaft erhöhen.

Der vierte Sektor, der in dieser Stellungnahme exemplarisch behandelt wurde, ist die öffentliche Verwaltung (vgl. Kapitel 8). Hier lässt sich in den vergangenen Jahren in vielen Ländern ein zunehmender Einsatz von Software zur Entscheidungsunterstützung in einer Vielzahl unterschiedlicher Bereiche beobachten, beispielsweise bei der Bewertung der Arbeitsmarktchancen Jobsuchender, der Prüfung und Vergabe von Sozialleistungen, bei Vorhersagen im Sozialwesen oder auch in diversen polizeilichen Einsatzkontexten. Auch wenn in diesen Fällen Entscheidungen nach wie vor von Menschen getroffen werden, so präkonfigurieren maschinelle Diagnosen und Prognosen diese doch in zunehmendem Ausmaß.

Häufig wird der Einsatz solcher Systeme einerseits mit einer Steigerung der Effizienz von Verwaltungsvorgängen begründet und andererseits mit einer möglichen Verbesserung der Qualität von – teils sehr tiefgreifenden – Entscheidungen, insbesondere bei komplexen Datenlagen. Demgegenüber stehen zahlreiche ethische Fragen und Probleme. Es ist nicht erwiesen, dass die Verwendung von Entscheidungssystemen zwangsläufig zu besseren Entscheidungen führt. Zudem

konnte in zahlreichen Studien gezeigt werden, dass insbesondere datenbasierte Systeme existierende gesellschaftliche Ungleichheiten oftmals reproduzieren. Durch den Einbau in scheinbar neutrale Entscheidungssysteme wird daher Diskriminierung zum einen perpetuiert und zum anderen unsichtbar gemacht. Gerade in Bezug auf Entscheidungen, die eine hohe Tragweite haben oder besonders vulnerable Gruppen betreffen, ist daher Sorge zu tragen, dass diese Systeme nicht nur genau sind, sondern auch offengelegt wird, mittels welcher Methoden mögliche Diskriminierungseffekte geprüft und mitigiert werden.

Umgekehrt kann der Einsatz von KI auch existierende Diskriminierungen in menschlicher Entscheidungspraxis identifizieren: Mit der Analyse früherer Entscheidungen wird es möglich, gesellschaftliche Diskriminierung aufzudecken, zu belegen und Änderungen einzufordern. Die Verbesserung von Entscheidungen, die Minimierung von Fehlern und das Verhindern bzw. Ausgleichen existierender Diskriminierung sind wichtige Ziele für den Einsatz von algorithmischen Systemen in der öffentlichen Verwaltung. Damit diese die Handlungsmöglichkeiten der verschiedenen Beteiligten jedoch erweitern und nicht vermindern, ist es entscheidend, dass Qualitätsstandards verbindlich festgelegt werden und deren Überprüfung durch externe Akteure möglich ist. Gerade in staatlichen Kontexten müssen hier höchste Anforderungen an Transparenz und Diskriminierungsminimierung angelegt werden.

10 ENTFALTUNG VON QUERSCHNITTSTHEMEN UND EMPFEHLUNGEN

Die Darstellung der soziotechnischen Entwicklungen und deren ethische Analyse in den vier Anwendungskontexten zeigen, dass es eine Reihe von Querschnittsthemen und -herausforderungen gibt, die sich – teils in unterschiedlichen Ausprägungsgraden und Formvarianten – durch alle vier Bereiche ziehen. Sie dürften auch in anderen Anwendungskontexten von „intelligenten" Maschinen eine Rolle spielen, die in dieser Stellungnahme nicht behandelt wurden. Um im Hinblick auf die Erweiterung menschlicher Handlungsfähigkeit und Autorschaft zukünftig einen guten gesellschaftlichen Umgang und eine entsprechende Gestaltung zu gewährleisten, können solche Querschnittsfragen nicht nur innerhalb der einzelnen Bereiche – sektoral – angegangen werden. Es werden darüber hinaus vernetzte, bereichsübergreifende Ansätze notwendig, um den im Folgenden skizzierten Themen und Fragen adäquat gerecht zu werden, die gleichzeitig ausreichende Feinkörnigkeit erlauben, um der ebenso erforderlichen Kontextsensitivität zu entsprechen. Solches gleichermaßen horizontale wie vertikale gestaltende Denken stellt eine Herausforderung insbesondere für die Politikgestaltung und die etwaige zukünftige Regulierung dar. Während dies bei einzelnen Themen bereits erkannt und – wenngleich nicht immer hinreichend – umgesetzt ist, beispielsweise in Bezug auf Probleme und Lösungen hinsichtlich eines verantwortlichen Umgangs mit Daten, bestehen mit Blick auf andere Themen noch erhebliche Defizite, beispielsweise in Bezug auf (Pfad-)Abhängigkeiten und Resilienz technologischer Infrastrukturen. Die folgende Darstellung von Querschnittsthemen und die Empfehlungen sollen daher als Anregung für eine breitere Debatte dienen, wie für zukünftige Politik- und Technikgestaltung gleichzeitig und im

Zusammenspiel mit sektoralen Aspekten immer auch übergreifende Fragen in den Blick genommen werden können und müssen.

10.1 Querschnittsthema 1: Erweiterung und Verminderung von Handlungsmöglichkeiten

Obwohl KI-Anwendungen in allen in dieser Stellungnahme erwähnten Bereichen neue Handlungsoptionen erschließen können, fallen die Potenziale für die tatsächliche Erweiterung menschlicher Handlungsmöglichkeiten bislang sehr unterschiedlich aus – nicht nur zwischen den einzelnen Feldern, sondern auch innerhalb verschiedener Segmente ein und desselben Bereichs.

Wie das Beispiel der Medizin zeigt, können insbesondere datenintensive Fächer wie die Onkologie von der Implementierung solcher Tools zur Verbesserung der Diagnostik und Therapie zum Wohle der Patientinnen und Patienten erheblich profitieren. Entscheidend ist dabei jedoch deren Einbettung in eine vertrauensvolle Arzt-Patienten-Beziehung, um der individuellen Besonderheit der jeweiligen Lebensumstände gerecht zu werden.

Ganz ähnlich ist auch der Einsatz von KI-Anwendungen im weiten Bereich der schulischen Bildung differenziert zu beurteilen. Während sich etwa auf dem Gebiet von Lehr- und Lernsystemen sowie intelligenten Tutorsystemen bereits gegenwärtig sinnvolle Einsatzgebiete beispielsweise zur Verbesserung individueller Lernprozesse abzeichnen, ist das Nutzen-Schaden-Potenzial bei anderen Anwendungen – wie zum Beispiel die Privatsphäre der Lernenden stark beeinträchtigenden Classroom Analytics – kritischer zu beurteilen.

Auch im Bereich der öffentlichen Kommunikation und Meinungsbildung ergibt sich ein breites Spektrum zwischen einer Erweiterung der Handlungsmöglichkeiten und ihrer

Verengung. Dies zeigt sich exemplarisch daran, dass die Möglichkeiten, sich breit und umfassend zu informieren und mit vielen Menschen schnell und kostengünstig zu kommunizieren, so groß sind wie nie zuvor. Im Gegenzug kann es jedoch auch zur Verminderung von Handlungsfähigkeit kommen, etwa wenn die freie Meinungsbildung gestört oder gar manipuliert wird. Auch wenn Existenz, Beschaffenheit und Ursachen von Effekten wie Filterblasen, Echokammern oder (Wahl-)Manipulation genauso kontrovers diskutiert werden wie mögliche Gegenmaßnahmen, so gilt es doch, Sorge zu tragen, dass Informations- und Kommunikationstechnologien so gestaltet werden, dass menschliche Autorschaft gestärkt und nicht vermindert wird.

In der öffentlichen Verwaltung werden die Auswirkungen von Technologien auf Handlungsfähigkeiten ebenso deutlich. Softwaresysteme, die beispielsweise Risikoscores erstellen, haben das Ziel, Personen dabei zu unterstützen, bessere Entscheidungen zu treffen. Idealerweise sollen solche Systeme die Handlungsfähigkeit also erweitern. In der Realität ist dies gleichwohl nicht immer der Fall: Zum einen kann durch blindes Befolgen algorithmischer Empfehlungen (Automation Bias) die Handlungsfähigkeit entscheidender Personen verringert werden. Zum anderen ist auch zu betrachten, inwiefern die Delegation von Entscheidungen an Softwaresysteme die Autorschaft und Handlungsmöglichkeit derer beeinträchtigt, die von den Entscheidungen betroffen sind.

Eine sektorübergreifende Gemeinsamkeit hinsichtlich der angestrebten Erweiterung menschlicher Handlungspotenziale besteht darin, dass die komplette Ersetzung menschlicher Akteure durch KI-Systeme sich überall dort verbietet, wo die konkrete zwischenmenschliche Begegnung eine notwendige Voraussetzung für die Erreichung der jeweiligen Handlungsziele darstellt. Darüber hinaus besteht die Notwendigkeit, die Differenzen eines KI-Einsatzes in den einzelnen Handlungsbereichen sorgfältig zu beachten. Diese betreffen nicht nur die unterschiedliche Sensibilität von Daten und ihre jeweilige

Bedeutung für das Leben und die Privatsphäre der Betroffenen, sondern auch den Bezug zu ganz verschiedenen Gütern (z. B. Gesundheit, Bildung, Teilhabe, Sicherheit).

Empfehlung

» *Empfehlung Querschnittsthema 1:* Da die Vor- und Nachteile von KI-Anwendungen für verschiedene Personengruppen sowie die Gefahr des Verlustes bestimmter Kompetenzen bei den Personen, die solche Systeme anwenden, erheblich variieren, bedarf es sowohl einer differenzierten Planung des KI-Einsatzes in unterschiedlichen Handlungsfeldern, welche die jeweiligen Zielsetzungen und Verantwortlichkeiten präzise benennt, als auch einer zeitnahen Evaluation der tatsächlichen Folgen eines solchen Einsatzes, um die Systeme besser an die spezifischen Handlungskontexte anzupassen und sie fortlaufend zu verbessern.

10.2 Querschnittsthema 2: Wissenserzeugung durch KI und Umgang mit KI-gestützten Voraussagen

Korrelationen und Datenmuster sind nicht mit Erklärungen und Begründungen von Ursachen von Ereignissen gleichzusetzen. Vielmehr müssen Daten nicht nur qualitativ evaluiert, sondern auch normativ beurteilt werden. Diese Beurteilung setzt die Fähigkeit voraus, Handlungen als Zweckrealisierungsversuche auszuführen, zu verstehen und zu überprüfen.

Mit dem Zuwachs qualitativ hochwertiger und relevanter Daten sowie verbesserter Methoden der Datenanalyse steigt häufig die Genauigkeit von Vorhersagen. Dennoch bestehen bei probabilistischen Methoden immer Restunsicherheiten. Dadurch ergibt sich die Frage, welcher Grad an epistemischer Sicherheit gesellschaftlich oder individuell für wünschenswert bzw. unerlässlich gehalten wird, um Handlungen und

Entscheidungen zu legitimieren. Auch wie Fehler erster oder zweiter Art, das heißt falsch-positive und falsch-negative Ergebnisse, und deren Folgen für unterschiedliche Betroffene zu bewerten sind, ist eine normative Frage, die den methodischen Entscheidungen in der Technikentwicklung vorausgehen muss.

Eine durch KI instrumentell unterstützte Daten- und Informationsbearbeitung sowie Wissensproduktion, die den Einfluss menschlicher Schwächen (z. B. Müdigkeit, Lustlosigkeit, Ehrgeiz, Bestechlichkeit) vermindert, würde epistemische und praktische Kompetenzen des Menschen grundsätzlich erweitern. Ein KI-gestützte Software allerdings, welche auf intransparenter Probabilitätsbasis praktische Kompetenz übernimmt, würde die menschliche Handlungssouveränität und Verantwortung hingegen vermindern. Wollte man versuchen, den Menschen als Handlungssouverän zu ersetzen, käme es zu einer Diffusion oder vollständigen Eliminierung von Verantwortung.

Für die ethische Beurteilung des Einsatzes von KI ist relevant, dass durch diesen in allen vier in dieser Stellungnahme angesprochenen Anwendungsbereichen erhebliche funktionale Verbesserungen erreicht wurden und weiterhin erwartbar sind. Dies gilt bei Diagnose und Therapie in der Medizin, im Unterricht in der schulischen Bildung, bei Information und Austausch von Argumenten im Bereich der öffentlichen Kommunikation und Meinungsbildung und schließlich in der öffentlichen Verwaltung. In allen Bereichen wird jedoch eine grundsätzlich normativ problematische Schwelle überschritten, wenn funktionale Verbesserungen (eventuell sogar unbemerkt) in eine Ersetzung moralischer Kompetenz und damit verbundener Verantwortung hinübergleiten.

Diese Schwelle ist beispielsweise im Gesundheitsbereich überschritten, wenn diagnostische und therapeutische Interventionen unter tatsächlicher Umgehung einer informierten Einwilligung der Patientin oder des Patienten (unter Umständen angetrieben durch Personalmangel) in Gänze digitalen

Systemen überlassen werden. Im Bereich der Bildung wäre dies der Fall, wenn nicht nur routinierte Lehrprozesse an Software delegiert, sondern Lehrpersonen vollständig technologisch ersetzt würden. Im Bereich der öffentlichen Kommunikation und Meinungsbildung wäre eine Grenze überschritten, wenn für die am öffentlichen Diskurs Teilnehmenden die moralische und rechtliche Verantwortung für Meinungen und Tatsachenbehauptungen nicht mehr erkennbar ist oder der Austausch von Argumenten durch digitale Maßnahmen praktisch aufgehalten wird. Im Bereich der öffentlichen Verwaltung gilt Entsprechendes, wenn Menschen nicht erkennen können, dass Verwaltungsentscheidungen durch digitale Agenten übernommen wurden oder eine solche Übernahme effektive Kontrollen verhindert. In allen Kontexten ist das entscheidende Kriterium für eine ethische Beurteilung die Möglichkeit der Identifizierung und der moralischen bzw. rechtlichen Inanspruchnahme von Agierenden, die in der Lage und auch verpflichtet sind, Handlungsverantwortung zu übernehmen.

Empfehlung

» *Empfehlung Querschnittsthema 2:* Der Einsatz KI-gestützter digitaler Techniken ist im Sinne der Entscheidungsunterstützung und nicht der Entscheidungsersetzung zu gestalten, um Diffusion von Verantwortung zu verhindern. Er darf nicht zulasten effektiver Kontrolloptionen gehen. Den von algorithmisch gestützten Entscheidungen Betroffenen ist insbesondere in Bereichen mit hoher Eingriffstiefe die Möglichkeit des Zugangs zu den Entscheidungsgrundlagen zu gewähren. Das setzt voraus, dass am Ende der technischen Prozeduren entscheidungsbefugte Personen sichtbar bleiben, die in der Lage und verpflichtet sind, Verantwortung zu übernehmen.

10.3 Querschnittsthema 3: Gefährdung des Individuums durch statistische Stratifizierung

Mit auf der Grundlage von Big Data erstellten Korrelationen werden menschliche Individuen Kohorten (Strata) zugeordnet (z. B. regional definierte Kohorten wie die Anwohnenden einer Straße oder Alterskohorten), die ohne jede soziale Erlebnisqualität sein können. Die Bildung solcher Kohorten und die auf ihrer Basis durch Algorithmen produzierten Voraussagen können für diejenigen, die solche Ansätze verwenden, durchaus nützlich sein, versprechen sie doch über die Gesamtmenge der Entscheidungen hinweg eine Erhöhung von Effektivität und Qualität. Probleme treten freilich für Individuen auf, die von solchen kollektiven Schlüssen betroffen sind – insbesondere dann, wenn die statistisch getroffene Diagnose oder Prognose in ihrem konkreten Fall nicht zutrifft.

Im Bereich der Medizin zeigt sich diese strukturelle Asymmetrie, wenn beispielsweise ein Individuum aufgrund diagnostizierter Merkmale einer Kohorte zugeordnet und aufgrund von Algorithmen eine therapeutische Strategie festgelegt wird. Man kann von einem „statistischen Kollektivismus" sprechen, gegenüber dem das Individuum geltend machen können muss, *als Individuum* betrachtet und behandelt zu werden. Das bedeutet, dass im Rahmen der Medizin als praktischer Wissenschaft eine korrelative Zuordnung eines Individuums zwar ein gutes diagnostisches Indiz sein kann, das wiederum eine gute praktische Heuristik rechtfertigt, aber grundsätzlich nicht die einzige Evidenzgrundlage für eine Behandlung sein darf. Epidemiologische Studien grenzen entsprechend denkbare prioritäre Diagnosen ein und legen auf dieser Grundlage bestimmte Therapien nahe. Sie sollten jedoch nie allein die Grundlage für eine Behandlung sein.

Die gleichen Probleme betreffen auch die datenbasierte Diagnostik und Prognose im Bereich der öffentlichen Verwaltung oder der schulischen Bildung. Auch wenn solche Systeme

das Ziel haben, nicht nur die Effektivität, sondern auch die Qualität von Entscheidungen zu verbessern, so bleibt ein unüberbrückbarer Spalt zwischen der je individuellen Lebenssituation und deren statistischer Annäherung. Auch wenn 99 von 100 Personen mit einem bestimmten Risikoprofil erneut straffällig werden oder keinen Schulabschluss erreichen werden, so wissen wir doch nicht, ob diese eine Person, über die zu entscheiden ist, nicht genau die eine Ausnahme ist.

Das Individuum entzieht sich einer vollständigen statistischen Annäherung. Jedoch wirkt jeder Versuch einer statistischen Annäherung möglicherweise auf das Individuum zurück. Genauer: Die Probleme der datenbasierten Voraussagen werden noch einmal um eine Dimension verschärft, wenn das Referenzobjekt, auf das sich Wahrscheinlichkeitsaussagen beziehen, durch diese Aussagen selbst beeinflussbar ist, wie das für alle Kontexte menschlichen Handelns, so zum Beispiel in der Volkswirtschaft oder im Gesundheitssystem, zutrifft. Dies gilt sowohl für Individuen als auch für Kollektive. Die Rückwirkungen von Klassifikationen und statistischen Vorhersagen insbesondere in Medizin, Sozialwissenschaften und Psychologie auf die betroffenen Individuen hat Ian Hacking mit dem Begriff „Loopingeffekt" beschrieben.[426] Aber auch auf kollektiver Ebene können beispielsweise aus epidemiologischen Studien gewonnene konditionierte Voraussagen bei passender öffentlicher Verbreitung kollektive Verhaltensweisen erheblich verändern – derart, dass die Datengrundlage nicht mehr verlässlich ist.

Diese Rückwirkung von Klassifikation und Prognostik auf Individuen wie Kollektive zeigt sich nicht zuletzt auch in Bezug auf soziale Medien im Bereich der öffentlichen Kommunikation und Meinungsbildung. Der Begriff des Hypernudge verweist hierbei auf die Besonderheit datenbasierter Systeme, die im Hintergrund und in Echtzeit die Informationsumgebung von Individuen, basierend auf statistischen

426 Hacking 1995.

Analysen, permanent anpassen und dadurch die weitere Informationsauswahl und die Entscheidungsmöglichkeiten präkonfigurieren.[427]

Empfehlung

» *Empfehlung Querschnittsthema 3:* Neben einer Analyse der konkreten und naheliegenden Probleme datenbasierter Software, beispielsweise in Bezug auf den Schutz der Privatsphäre oder die Verhinderung von Diskriminierung, gilt es, auch die langfristigen Auswirkungen dieser statistischen Präkonfiguration von Individuen sowie deren Rückwirkung – im Sinne einer Erweiterung oder Verminderung der Handlungsmöglichkeiten – auf Individuen wie Kollektive für alle Sektoren sorgfältig zu beleuchten. Darüber hinaus gilt, dass Einzelfallbeurteilungen grundsätzlich wichtig bleiben. KI-basierte Beurteilungen und Vorhersagen können unter günstigen Bedingungen ein Hilfsmittel sein, aber kein geeignetes Instrument der *definitiven* Lagebeurteilung und Entscheidung. Pragmatische und heuristische Faktoren wie die Prüfung der Kohärenz mit anderen Evidenzquellen oder Erfolgseinschätzungen spielen eine nicht zu vernachlässigende Rolle.

10.4 Querschnittsthema 4: Auswirkungen von KI auf menschliche Kompetenzen und Fertigkeiten

Der Einsatz von KI-Anwendungen in ganz unterschiedlichen Lebensbereichen beeinflusst menschliche Handlungsfähigkeit und hat erheblichen Einfluss auf den Erwerb und den Erhalt menschlicher Kompetenzen und Fertigkeiten.

So zeigt sich im Bereich der Medizin, dass der fachgerechte Einsatz von KI-Anwendungen dazu geeignet ist, diagnostische

427 Yeung 2017.

Kompetenzen zu verbessern und zu einer deutlichen Reduktion von Fehldiagnosen führen kann. Auch in der schulischen Bildung dienen digitale Technologien in vielfältiger Weise der Kompetenzerweiterung. Im Bereich der öffentlichen Kommunikation und Meinungsbildung erweitern digitale Technologien die informationellen und kommunikativen Möglichkeiten der Beteiligten und erlauben damit eine eventuell kompetentere Teilnahme am öffentlichen Diskurs. In der öffentlichen Verwaltung wiederum sollen Softwaresysteme die Kompetenz der Beschäftigten erhöhen, gute und angemessene Entscheidungen zu treffen, etwa beim Einsatz von Risikobewertungen zur Prognose und Prävention von Gefahrenlagen.

Allerdings hat der Einsatz von KI-Anwendungen nicht nur das Potenzial, Kompetenzen zu erweitern, sondern auch, diese zu vermindern. Dies kann wiederum mit einem Beispiel aus dem medizinischen Bereich veranschaulicht werden: Eine Software, die Röntgenbilder präzise nach Auffälligkeiten untersucht und dabei statistisch deutlich besser abschneidet als der Mensch, erweist sich als signifikante Erweiterung bisheriger menschlicher Handlungsmöglichkeiten. Zugleich geht mit dem Einsatz des Systems das Risiko einher, dass der die Anwendung nutzende Mensch die eigenen Fähigkeiten einbüßt, Röntgenbilder unter diesem Gesichtspunkt zu analysieren. Dieser Effekt des Deskillings kann etwa deshalb eintreten, weil die jeweilige Fähigkeit zu ihrem Erhalt der steten Einübung bedarf, sie nunmehr aber nicht hinreichend häufig praktiziert wird, da die Technologie den Menschen mehr und mehr unterstützt bzw. seine Aufgabe vornehmlich übernimmt.[428] Ein weiterer Grund für Deskilling kann das große Vertrauen in eine nahezu perfekte Technologie bei gleichzeitigem Misstrauen in die eigenen Fähigkeiten sein (Automation Bias). Der einzelne Mensch könnte geneigt sein, sich weniger auf das eigene als auf das technisch erzeugte Urteil zu verlassen – mit der Folge, dass wiederum Fähigkeiten nicht eingeübt werden und

428 Bainbridge 1983.

so mit der Zeit verloren gehen. Angesichts des verstärkten Einsatzes von KI-Anwendungen wird Deskilling in verschiedenen gesellschaftlichen Bereichen zunehmend befürchtet. Freilich ist Kompetenzverlust als Effekt nicht allein auf digitale oder gar KI-basierte Technologien begrenzt, sondern kann letztlich bei nahezu allen Werkzeugen beobachtet werden, derer Menschen sich zur Vereinfachung ihrer Aufgaben bedienen.

Den offenkundigen Vorteilen der Erweiterung von Fähigkeiten und Handlungsmöglichkeiten durch KI-basierte Systeme stünden also etwaige Nachteile gegenüber, die entstehen, wenn es zu einer Verringerung von Fähigkeiten und Fertigkeiten menschlicher Akteure in bestimmten gesellschaftlichen Bereichen käme. Wann dies in der Gesamtschau akzeptabel und wann problematisch wäre, lässt sich nicht pauschal oder übergreifend beantworten, sondern muss kontext- und fallspezifisch bestimmt werden. Weil die Nutzung von KI-Anwendungen dazu führen kann, dass bedeutsame menschliche Fähigkeiten nachlassen bzw. ganz verkümmern, können Abhängigkeiten von diesen Technologien entstehen. Es ist jedoch gerade angesichts der enormen Zunahme digitaler Technologien nicht ohne Weiteres ausgemacht, dass die Voraussetzungen für deren Einsatz, wie zum Beispiel eine stabile Stromversorgung, immer sicher gewährleistet werden können. Dies kann menschliche Fähigkeiten mitunter kurzfristig wieder bedeutsam werden lassen. Das spezifische Risiko von Kompetenzverlusten im Zusammenhang mit dem Einsatz von KI-Anwendungen liegt demnach in der Besonderheit der Tätigkeiten, die der Technik überlassen werden – handelt es sich dabei um gesellschaftlich besonders bedeutsame oder kritische Einsatzbereiche, ist ein Verlust von menschlichen Kompetenzen und Fertigkeiten ein ernstzunehmendes Risiko.

Allerdings ließe sich einem solchen Risiko entgegenwirken. Droht Kompetenzverlust und ist dieser gesellschaftlich unerwünscht, bedarf es dafür effektiver Vorkehrungen zur Vermeidung dieses Effekts. Eine Lösung kann beispielsweise darin liegen, dass die Technologie nur dann zum Einsatz gebracht

werden darf, wenn sichergestellt ist, dass die Betroffenen die Fertigkeiten weiter regelmäßig trainieren. In der Medizin etwa könnte dies neben noch zu entwickelnden spezifischen Fortbildungsprogrammen bedeuten, Befundungen nie ganz an einen Algorithmus abzugeben.

Gleichwohl bleibt auch unter dieser Voraussetzung der Einwand bestehen, dass infolge eines immer durchgreifenderen Einsatzes von KI-Anwendungen in unterschiedlichen Lebensbereichen Menschen mehr und mehr dazu verleitet werden könnten, eigene Aufgaben an die Technik zu delegieren, weil diese als (vermeintlich) überlegen angesehen wird. Hieraus ergibt sich ein Risiko für die individuelle Selbstwahrnehmung. Mit zunehmender Delegation von Aufgaben an Technologien kann der Eindruck entstehen, immer mehr Kontrolle über wesentliche Bereiche des eigenen Lebens abzugeben. Im Extremfall könnte eine regelmäßige Delegation von Entscheidungen einen Effekt auf die Wahrnehmung des Selbst als Autor des eigenen Geschickes haben und sogar bürgerschaftliches Engagement reduzieren. Vor allem aber könnten „Kipppunkte" einer solchen passiven Selbstwahrnehmung schleichend erreicht und überschritten werden, ohne dass noch effektiv interveniert werden könnte.

Angesichts dieser Überlegungen sollten das Maß an Delegation, etwaige Verlagerungen von Aufgaben aus einem Lebensbereich in andere sowie die allgemeine gesellschaftliche Entwicklung infolge des verstärkten Einsatzes von KI-Anwendungen in dieser Hinsicht sorgsam beobachtet werden. Wenn das beschriebene Risiko jedoch mehr als eine denkbare Möglichkeit darstellt, sich also zu einem relevanten Faktor verdichtet, erscheinen Gegensteuerungsmaßnahmen erforderlich.

Empfehlung

» *Empfehlung Querschnittsthema 4:* Ob und inwiefern beim Einsatz von KI-Anwendungen Verluste menschlicher Kompetenz auftreten, die als unerwünscht eingestuft werden, muss sorgfältig beobachtet werden. Bei der

Entwicklung und dem Einsatz neuer Technologien sind solch unerwünschte Kompetenzverluste durch eine sinnvolle Gestaltung des Zusammenspiels von Mensch und Technik, durch angemessene institutionelle und organisatorische Rahmenbedingungen sowie durch gezielte Gegenmaßnahmen wie etwa spezifische Trainingsprogramme zu minimieren bzw. zu kompensieren. Kompetenzverluste können sowohl individueller als auch kollektiver Natur sein. So gilt es zu verhindern, dass die Delegation von Aufgaben an Technologien dazu führt, dass Gesellschaften übermäßig anfällig werden, wenn diese Technologien (zeitweise) ausfallen. Jenseits dieser systemischen Aspekte müssen negative Auswirkungen solcher Delegation auf die individuelle Autonomie oder Selbstwahrnehmung mitigiert werden.

10.5 Querschnittsthema 5: Schutz von Privatsphäre und Autonomie versus Gefahren durch Überwachung und Chilling-Effekte

Ein weiteres wichtiges übergreifendes Thema, welches sich durch alle Anwendungsbereiche zieht, ist die Gefahr der Überwachung und die Möglichkeit auftretender *Chilling*-Effekte, die nicht nur die Privatsphäre von Menschen, sondern auch deren Autonomie gefährden.[429] Viele der zuvor behandelten Technologien sind auf die Sammlung und Auswertung großer Mengen an personenbezogenen Daten angewiesen. Die Erfassung solcher Daten und die Möglichkeit, auf ihrer Basis sensible Prognosen zu erstellen, beeinträchtigen nicht nur die Privatsphäre der Personen, von denen die Daten stammen, sondern machen sie auch vulnerabel gegenüber möglichen

429 Penney 2017; Solove 2006.

Benachteiligungen oder Manipulationen, die aus der Verarbeitung der Daten resultieren können.

Aber auch wenn Daten nicht personenbezogen ausgewertet, das heißt keine Rückschlüsse auf das Individuum gezogen werden, kann sich bereits die Sorge vor der Möglichkeit solcher Rückschlüsse negativ auf die Autonomie und Entfaltungsfreiheit von Menschen auswirken. Der Chilling-Effekt beschreibt in diesem Kontext Rückwirkungen auf Menschen, die Sorge haben, dass ihr Verhalten beobachtet, aufgezeichnet oder ausgewertet wird. Besonders sichtbar werden solche Chilling-Effekte im Kontext von öffentlicher Kommunikation und Meinungsbildung. So können Menschen davor zurückschrecken, nach relevanten, aber möglicherweise sensiblen Themen im Internet zu suchen oder als kontrovers wahrgenommene Inhalte zu lesen, in dem Wissen, dass ihr Onlineverhalten getrackt wird, und aus der Sorge heraus, in welcher Art und Weise sich das möglicherweise nachteilig auf sie auswirken könnte.[430] Es geht hier also einerseits um die tatsächlichen negativen Effekte der Überwachung von Individuen und andererseits um Effekte, die als Anpassung an die Sorge vor dieser Überwachung entstehen.

Invasives Tracking führt nicht nur dazu, dass Menschen personalisierte, passgenaue Werbung zugespielt wird, sondern auch, dass Datenspuren, die sie online hinterlassen, im Aggregat und Zusammenspiel mit weiteren Datenspuren (sowohl aus anderen Quellen als auch von anderen Personen) etwa mittels Datenbroker zu vielfältigen, für die betroffene Person nicht nachvollziehbaren Zwecken zweitverwertet werden können. Das Geschäftsmodell von Datenbrokern stellt eine Beeinträchtigung der Privatsphäre dar und beruht auf der Fiktion, dass eine Zustimmung zu allgemeinen Geschäftsbedingungen von Onlineangeboten einer informierten Einwilligung gleichkommt.

430 Büchi/Festic/Latzer 2022.

Chilling-Effekte wiederum sind das Resultat des Wissens um oder auch nur der Ahnung von solchen Praktiken. Das eigene Verhalten verändert sich so, als würde man tatsächlich überwacht. Nicht nur der Informationskonsum wird einer Art Selbstzensur unterworfen, sondern auch die freie Meinungsäußerung kann online beeinträchtigt werden. Das Tracking insbesondere des Medienverhaltens von Individuen beeinträchtigt somit nicht nur deren intellektuelle Privatsphäre, sondern hat möglicherweise auch negative Auswirkungen auf die Kommunikationsfreiheit, auf individuelle Kreativität und kann somit der Persönlichkeitsentwicklung und der demokratischen Selbstverständigung abträglich sein.

Ähnliche Probleme können auch in gesundheitlichen Kontexten beziehungsweise an der Schnittstelle zwischen Onlineverhalten und Medizin entstehen. Wenn Onlineverhalten getrackt wird und Datenspuren von Brokern weiterverkauft werden, so kann die Sorge, aufgrund der möglichen Diagnose oder Prognose etwa einer psychischen Erkrankung keine private Krankenversicherung mehr zu bekommen, dazu führen, dass nicht nach Informationen gesucht wird, die etwa Depression oder Suchtverhalten nahelegen könnten. Auch bei der Verwendung von Technologien im Gesundheitswesen selbst, beispielsweise beim Einsatz von datenbasierten Systemen in der Diagnostik oder auch bei der Kommunikation mit therapeutischen Chatbots, können Menschen, die solche Angebote nutzen, annehmen – ob berechtigt oder nicht –, dass sensible Daten oder Prognosen in die Hände unbefugter Dritter geraten können, und falsche Angaben machen, die möglicherweise zu falschen Diagnosen führen. Um etwaigen Chilling-Effekten entgegenzuwirken, gilt es also, unzulässige Datenweitergaben zu verhindern, insbesondere dann, wenn diese nicht im Interesse der Personen sind, von denen die Daten stammen, oder gar fremden, nicht gesundheitsbezogenen Zwecken dienen.

Auch in der schulischen Bildung müssen Gefahren von Überwachung und mögliche Chilling-Effekte ernst genommen werden. Dies gilt insbesondere in Bezug auf Technologien,

welche auf Audio- oder Videoüberwachung des Klassenraums basieren. Emotionserkennung oder Aufmerksamkeitsmonitoring kann nicht nur die Privatsphäre der Lernenden sowie gegebenenfalls der Lehrkräfte auf unzulässige Art und Weise beeinträchtigen. Bereits die Sorge vor einer solchen Überwachung, selbst wenn die Daten nicht aufbewahrt und nicht personenbezogen analysiert und ausgewertet würden, hat unter Umständen Auswirkungen auf das Verhalten der Beteiligten.

Die parallelen Gefahren von Überwachung und Chilling-Effekten stellen sich schließlich auch im Bereich der öffentlichen Verwaltung. Paradigmatisch hierfür sind Maßnahmen im Kontext von Predictive Policing wie beispielsweise der Chatkontrolle. Eine anlasslose Überwachung der Telekommunikation ist grundrechtlich inakzeptabel; bereits die Sorge vor einer solchen Überwachung könnte dazu führen, dass Menschen sich selbst zensieren.

Empfehlung

» *Empfehlung Querschnittsthema 5:* Die beschriebenen Phänomene sollten in ihrer Entstehung, Ausprägung und Entwicklung umfassend empirisch untersucht werden. Um sowohl dem Problem der Überwachung sowie den parallelen Gefahren durch etwaige Chilling-Effekte Rechnung zu tragen, müssen angemessene und effektive rechtliche und technische (z. B. Privacy by Design) Vorkehrungen getroffen werden, die dem übermäßigen Tracking von Onlineverhalten und dem Handel mit personenbeziehbaren Daten Einhalt gebieten. Die Interessen der Datensubjekte müssen hierbei im Mittelpunkt stehen. Insbesondere ist dabei auf besonders vulnerable Gruppen zu achten, da viele der Einsatzkontexte zudem von asymmetrischen Machtverhältnissen gekennzeichnet sind. Es muss Sorge getragen werden, dass die Erweiterung der Handlungsmöglichkeiten einiger nicht zulasten der Verminderung der Handlungsmöglichkeiten anderer, insbesondere benachteiligter Gruppen stattfindet.

10.6 Querschnittsthema 6: Datensouveränität und gemeinwohlorientierte Datennutzung

Wie zuvor erläutert, bergen viele der in dieser Stellungnahme vorgestellten Technologien insbesondere aufgrund der Nutzung vielfältiger Daten zwar zahlreiche Risiken, sie bringen aber gleichzeitig große Chancen mit sich, auf die wir als Gesellschaft ungern verzichten würden. Es müssen also Lösungen entwickelt werden, wie Daten sinnvoll für verschiedene wichtige Zwecke genutzt werden können, ohne zugleich den Schutz der Privatsphäre der Datensubjekte unzulässig zu beeinträchtigen.[431]

In diesem Zusammenhang stellt sich die Frage, ob das derzeitige Datenschutzrecht bzw. die herrschende Datenschutzpraxis diesen beiden Zielen gerecht wird. Hier kommt Kritik aus beiden Richtungen: Einerseits wird die Privatsphäre von datenliefernden Personen gerade im Kontext von sozialen Medien und der zugrundeliegenden Datenökonomie nicht ausreichend geschützt. Andererseits wirft die aktuelle Datenschutzpraxis gerade in der öffentlichen Forschung vielfach Probleme auf – auch und gerade dort, wo es explizit um gemeinwohlorientierte Aktivitäten und eben nicht um (unbemerkte) Nachverfolgung geht, die die Privatsphäre und Interessen der Personen, die ihre Daten zur Verfügung stellen, verletzt. Während also in manchen Handlungsfeldern berechtigte Sorgen

431 Diesem Zielkonflikt und den damit verbundenen, vielfältigen Herausforderungen hat sich bereits die Stellungnahme des Deutschen Ethikrates zu Big Data und Gesundheit aus dem Jahr 2017 gewidmet. Dort hat er die Notwendigkeit, individuelle Interessen und die Privatsphäre effektiv zu schützen, systematische Schadenspotenziale zu erkennen und zu verringern sowie gleichzeitig gemeinwohlorientierte Aktivitäten in der Datennutzung zu ermöglichen, im normativen Konzept der Datensouveränität zusammengeführt. Der Rat hat in der Stellungnahme eine Vielzahl von Empfehlungen vorgelegt, wie der Schutz der Privatsphäre unter den Bedingungen datengesättigter Lebenswelten verbessert werden kann; diese sind vielfach auch auf die hier verhandelten Anwendungsbereiche übertragbar (Deutscher Ethikrat 2017).

vor unbemerkten und weitreichenden Verletzungen von Privatsphäre und informationeller Selbstbestimmung herrschen, werden in anderen Kontexten durch strenge Auslegungen von Datenschutzregeln wichtige soziale Güter, etwa mit Blick auf Patientenversorgung und wissenschaftlichen Erkenntnisgewinn, aber auch die kommunale Daseinsvorsorge nicht oder nur sehr schwer erreicht.

Eine wesentliche Ursache für beide Probleme sind Charakteristika des Datenschutzrechts, die in der Vergangenheit umfassend kritisiert worden sind und die sich, verknappt gesprochen, aus dem zugrundeliegenden Individualismus und der damit verbundenen, überbetonten Rolle des Instruments der individuellen informierten Einwilligung zur Datennutzung ergeben.

Das geltende Datenschutzrecht ist auf die aktuellen Herausforderungen von KI nicht optimal vorbereitet. Durch den Fokus auf das initial datengebende Individuum werden systemische Risiken und Effekte auf andere Personen unzureichend berücksichtigt. Sogenannte relationale Schäden, die durch die Preisgabe der Daten anderer erfolgen können, können oft nicht hinreichend erfasst werden, etwa in Bezug auf Gruppenprivatsphäre oder auch Diskriminierung. Denn die angenommene enge Zweckbindung von Daten, die dieser Vorstellung zugrunde liegt, lässt sich allenfalls für die Datennutzung der ersten Instanz realisieren. Jede Weiterverwendung führt zu einer De- und Rekontextualisierung, zu der jeweils neue Zwecksetzungen gehören. Die Kaskade der Weiterverwendungen führt dabei sehr schnell aus dem Übersichtsbereich der Person, von der die Daten stammen, heraus. Zudem erschwert die stark auf das Individuum enggeführte Perspektive, dass systematische Chancen von Datennutzungen nicht bzw. unzureichend genutzt werden.

Das weiterhin datenschutzrechtlich zentrale Instrument der informierten Einwilligung, das sich aus der individualistischen Perspektive auf Datennutzung ergibt, führt wiederum in verschiedenen Kontexten ins Leere bzw. in die praktische

Dysfunktionalität: Insbesondere bei den beschriebenen Beispielen des Datentrackings im Bereich des medizinischen Onlineverhaltens, zum Beispiel bei Gesundheits-Apps oder in sozialen Medien allgemein, kann regelmäßig weder von Informiertheit noch von einer Einwilligung der Personen, die diese Angebote verwenden, ausgegangen werden. Dies hat sich auch im Gefolge der Einführung der Datenschutz-Grundverordnung nicht verändert; die entsprechenden Details der Einwilligungserklärungen werden bekanntermaßen oft nicht gelesen bzw. nicht inhaltlich wahrgenommen. Dort wiederum, wo explizite informierte Einwilligungsformate unabdingbar und auch breit in der Praxis eingeführt sind – etwa im Bereich der medizinischen Forschung –, können sie offenkundig sinnvolle Anwendungen stark erschweren oder verunmöglichen, weil sie, dem Primat der individuell orientierten engen Zweckbindung folgend, alle weiteren Datennutzungen verunmöglichen oder jedenfalls so ausgelegt werden. So wird etwa gerade in der öffentlich geförderten universitären Forschung durch eine restriktive Auslegungspraxis mit Blick auf Zweckbindung und informierte Einwilligungsformate die Sekundärnutzung von Patientendaten stark erschwert, selbst wenn Patientinnen und Patienten initial in eine intensive Nutzung ihrer Daten eingewilligt haben; der Spielraum, den die Datenschutz-Grundverordnung hier bietet, wird oft nicht genutzt.

Dabei, so hat es der Deutsche Ethikrat bereits 2017 in seiner Stellungnahme zu Big Data und Gesundheit unterstrichen, sind aus ethischer Sicht nicht nur Präferenzen der individuellen Binnensphäre, sondern auch Gesichtspunkte der Solidarität im Umgang mit Datensammlung und -nutzung zu bedenken, die eine besondere Aktualität bei KI-Anwendungen erlangen. In der gängigen Metapher gesprochen, sind Daten grundsätzlich ein wertvoller Rohstoff, den die einzelne Person unter bestimmten Umständen dem Gemeinwesen zur Verfügung stellen sollte. Die Einordnung von Daten als potenzielles Gemeingut steht nicht in striktem Widerspruch zum individuellen Anspruch auf Schutz der Privatsphäre. Vielmehr

widerstreitet sie dessen Absolutsetzung und pauschaler Vorrangstellung und verweist darauf, dass beide Vorgaben – kontextspezifisch – in einen angemessenen Ausgleich zu bringen sind. Daher sollte die Möglichkeit bestehen, dass individuelle Akteure selbstbestimmt Daten im Interesse der Allgemeinheit zur Verfügung zu stellen, beispielsweise für die medizinbezogene Grundlagenforschung. Dem entspricht der Fokus im unionalen Data Governance Act auf „Datenaltruismus".

Diese Spannungen sind wohlbekannt im Bereich der Medizin. Dort ist das Ungleichgewicht zwischen – vereinfacht gesprochen – „Trackingwildwuchs" etwa im App-Bereich auf der einen Seite und sehr restriktiver Praxis mit Blick auf klinische Sekundärnutzung von Daten für sinnvolle öffentliche und gemeinwohlorientierte Forschung auf der anderen Seite besonders ausgeprägt sowie bekannt.[432] Sie haben aber ebenso starke Relevanz in den anderen Anwendungsfeldern. Offenkundig ist dies auch bei datenintensiven Anwendungen in der schulischen Bildung der Fall, wo der Schutz der Privatsphäre sowohl von Lernenden als auch von Lehrkräften essenziell ist und immer gewährleistet sein muss, etwa durch entsprechende technische Voreinstellungen (Privacy by Design). Gleichwohl ist auch hier offenkundig, dass sich aus einer stärker gemeinwohlorientierten Perspektive vielfältige wichtige Nutzungen für derart gesammelte Daten ableiten lassen. So etwa wäre die verantwortliche Auswertung von Daten aus Tutorsystemen geeignet, um Indikatoren für edukatorische Ungleichheiten oder Nachteile im Bildungssystem für bestimmte Gruppen von Lernenden systematischer, schneller und vorausschauender zu erfassen und solche Probleme zielgenauer anzugehen. Ähnliche Konstellationen ergeben sich auch in der öffentlichen Verwaltung. Dort kann eine präzisere und stärker vorausschauende Erfassung etwa von Risikofaktoren für eine Kindeswohlgefährdung oder in der Bewährungshilfe wichtigen individuellen

432 Sachverständigenrat zur Begutachtung der Entwicklung im Gesundheitswesen 2021.

und gesellschaftlichen Zielen dienen. Zugleich erfordert dies den hohen Schutz der Privatsphäre aller Betroffenen, um unzulässige Vorverurteilungen oder Stigmatisierungen zu vermeiden. Zusammenfassend scheinen sowohl die rechtlichen Rahmenbedingungen als auch die konkrete Anwendungspraxis in bestimmten Bereichen, insbesondere in den für die öffentliche Kommunikation und Meinungsbildung so zentralen sozialen Medien und der zugrundeliegenden Datenökonomie, nicht nur die Privatsphäre der Nutzenden unzureichend zu schützen. Auch eine mögliche Datensouveränität wird ad absurdum geführt, da vielfach weder das Kriterium der Informiertheit noch jenes der Einwilligung gewährleistet ist. Dies gilt vielfach auch für das Datensammeln und -verarbeiten im Hintergrund kommerzieller Technologien, die im Bildungswesen, der öffentlichen Verwaltung und im Medizinbereich eingesetzt werden.

Demgegenüber werden wichtige Chancen insbesondere im Bereich der öffentlichen Forschung durch eine teils übermäßig rigide Datenschutzpraxis vergeben, die etwa Fortschritten bei der Bildungsgerechtigkeit, der Früherkennung von Erkrankungen oder der flächendeckenden Vermeidung von Medikamentenverschreibungsfehlern entgegensteht sowie optimierter und zugleich maßvoller und verantwortlicher Risikoabschätzung im Sozialwesen zuwiderläuft.

Empfehlung

» *Empfehlung Querschnittsthema 6:* Mit Blick auf KI-Anwendungen müssen neue Wege gefunden werden, um innerhalb der jeweiligen Kontexte und bezüglich der jeweils spezifischen Herausforderungen und Nutzenpotenziale die gemeinwohlorientierte Daten(sekundär)nutzung zu vereinfachen bzw. zu ermöglichen und damit die Handlungsoptionen auf diesem Gebiet zu erweitern. Zugleich ist es essenziell, einen Bewusstseinswandel sowohl in der Öffentlichkeit als auch bei den praktisch tätigen Personen, die Datennutzung gestalten, herbeizuführen – weg von einer

vornehmlich individualistisch geprägten und damit verkürzten Perspektive hin zu einer Haltung, die auch systematische und gemeinwohlbasierte Überlegungen mit einbezieht und in einen Ausgleich bringt. Eine solche Haltung ist auch für die zukünftige Politikgestaltung und Regulierung deutlich stärker als bisher zugrunde zu legen. Nur so kann es gelingen, neben den Risiken, die sich aus breiterer KI-Anwendung ohne Zweifel ergeben, zugleich die wichtigen Chancen einer verantwortlichen Nutzung nicht aus dem Blick zu verlieren.

10.7 Querschnittsthema 7: Kritische Infrastrukturen, Abhängigkeiten und Resilienz

Infrastrukturen sind zentrale Elemente moderner Gesellschaften, die für das Funktionieren gesellschaftlicher Teilbereiche essenziell sind und deren Ausfall mit großen gesamtwirtschaftlichen Schäden verbunden ist. Im Zuge der Digitalisierung werden Infrastrukturen wie beispielsweise Stromnetze zunehmend digital überwacht und über das Internet gesteuert. Auf der anderen Seite werden digitale Technologien selbst zu Infrastrukturen. Dies gilt insbesondere für digitale Medien im Kontext der öffentlichen Kommunikation und Meinungsbildung: Kommerzielle Plattformen stellen hier zunehmend Infrastrukturen für den öffentlichen Diskurs dar.

Im Kontext von öffentlicher Verwaltung und Bildung, aber auch in Teilen der Medizin, stellt sich die infrastrukturelle Bedeutung digitaler Technologien teilweise ähnlich, teilweise sehr unterschiedlich dar. Die Gemeinsamkeit besteht darin, dass es in Teilen dieselben Unternehmen sind, die nicht nur die großen Plattformen betreiben, sondern auch Technologien für die Bereiche öffentliche Verwaltung, Bildung und Medizin anbieten. Der Unterscheid besteht darin, dass sich die Vulnerabilität der öffentlichen Verwaltung und Schulen und

teilweise auch der medizinischen Versorgung beispielsweise in der Coronapandemie gerade in einer unzureichenden Digitalisierung gezeigt hat.

Diese doppelte infrastrukturelle Bedeutung digitaler Technologien gilt es also zu berücksichtigen und unter ethischen Aspekten zu beleuchten. Hier ist auf folgende mögliche Problembereiche hinzuweisen:

Während Infrastrukturen von Menschen aufgebaut werden, um bestimmten Zwecken zu dienen, Menschen also die Systeme gestalten, hat der infrastrukturelle Charakter zur Folge, dass sich dieses Verhältnis allmählich verschiebt. Denn am Vorhandensein und Funktionieren der betreffenden Infrastruktur richten Menschen ihr Handeln aus. Im Zuge dieser sozialen Aneignung einer Infrastruktur entstehen Abhängigkeiten, die die menschliche Autonomie gefährden können. Dies zeigt sich beispielsweise im Kontext der Übernahme von Twitter durch Elon Musk. Dabei wurde deutlich, dass es keine vergleichbaren Alternativen gibt, auf welche im Falle grundlegender Änderungen der Geschäftsmodelle oder Moderationspraktiken ausgewichen werden könnte, ohne dass zentrale Aspekte der Plattformnutzung wegfallen würden.[433]

Die vorstehend geschilderte Problematik gilt für Infrastrukturen generell und ist nicht spezifisch auf KI bezogen. Wenn jedoch KI-gestützte ADM-Systeme zusehends in die Steuerung der Infrastrukturen integriert werden, kommt eine neue Form von Abhängigkeit hinzu. KI-Systeme sind nicht vollständig transparent und nachvollziehbar. Die infrastrukturelle Abhängigkeit würde also auch die Abhängigkeit von Systemen beinhalten, die zumindest zum Teil Blackboxes sind (vgl. Abschnitt 10.10). Im Zusammenhang mit dem Phänomen menschlicher Gewöhnung an entsprechende Systeme kann es zur Verfestigung von Verhaltensmustern in soziotechnischen

433 Zwar migrieren derzeit viele Twitter-Nutzerinnen und -Nutzer zu anderen Services wie beispielsweise Mastodon. Deren Funktionalität und Reichweite sind jedoch nicht mit Twitter deckungsgleich.

Konstellationen kommen, die eine kritische Reflexion und ein Hinterfragen der teils intransparenten ADM-Systeme erschweren können. Durch solche Pfadabhängigkeiten (vgl. Abschnitt 10.8) wird eine Eigendynamik eingeleitet, die den Wechsel auf andere Formen erschwert oder unmöglich macht und damit Optionen verbaut. Auch hier kann Twitter zur Illustration dienen: Sein Mehrwert besteht gerade in der Interaktion mit anderen Nutzenden; durch einen Wechsel auf ein anderes, gegebenenfalls dezentrales Medium wie Mastodon geht ein Teil des Nutzens jedoch verloren. Solche sogenannten *Netzwerkeffekte* erschweren den Wechsel zwischen Plattformen und erhöhen die Abhängigkeiten.

Die Abhängigkeit vom Funktionieren der digitalen Systeme betrifft längst nicht mehr nur die Informations- und Kommunikationsinfrastruktur. Dadurch, dass mittlerweile viele Infrastrukturen digitalisiert sind und über das Internet gesteuert werden, sind sie mit dem Internet als Quasi-Nervensystem zu einer Mega-Infrastruktur verbunden. Hackerangriffe, technisches Systemversagen oder soziotechnische Systemrisiken können auf diese Weise erheblich größere Reichweite haben als in separierten Infrastrukturen. Durch die fortwährende Komplexitätssteigerung der Infrastruktursysteme und ihre Steuerung steigt die gesellschaftliche und institutionelle Vulnerabilität weiter an. Demgegenüber ist die allgemeine Erfahrung, dass die Infrastrukturen zuverlässig funktionieren und Störungen vorübergehend sind. Dieses Funktionieren, so etwa die hohe Zuverlässigkeit der Stromversorgung und des Internets in westlichen Ländern, kann blind für die zunehmende Abhängigkeit von digitaler Technik machen. Schwere Wirtschaftskrisen, technische Systemeffekte, ein Kollaps der staatlichen Ordnung oder Hacker-Angriffe sind jedoch nicht unmöglich. Ethisches Vorsorgedenken gebietet, solche Abhängigkeiten zumindest bewusst zu machen und Strategien für digitale Blackouts zu entwickeln. Dies gilt umso mehr für intransparente, KI-getriebene Systeme, in denen sowohl Fehler als auch Angriffe noch schwerer zu entdecken und nachzuweisen sind.

Empfehlung

» *Empfehlung Querschnittsthema 7:* Um die Autorschaft menschlicher Akteure und deren Handlungsmöglichkeiten zu erweitern, müssen die Resilienz soziotechnischer Infrastrukturen gestärkt und die Abhängigkeit von individuellen Akteuren und Systemen minimiert werden. Dies umfasst zunächst die Notwendigkeit, die infrastrukturelle Bedeutung digitaler Technologien anzuerkennen und infolgedessen dem Schutz und der Resilienz kritischer digitaler Infrastrukturen mehr Aufmerksamkeit zuteilwerden zu lassen, auch im politischen Handeln. In allen Sektoren gilt es, einseitige Abhängigkeiten zu vermeiden, welche im Krisenfalle verletzlich und angreifbar machen. Für Nutzerinnen und Nutzer erfordert eine Verringerung der Abhängigkeit die Möglichkeit, zwischen Alternativen zu wählen, ohne große Teile der Funktionalität einzubüßen. Dies umfasst die Notwendigkeit von Interoperabilität, um einfach zwischen Systemen wechseln zu können. Hierfür ist auch der Auf- und Ausbau alternativer Infrastrukturen von besonderer Bedeutung. Im Kontext der öffentlichen Meinungsbildung erscheint die Etablierung unabhängiger, öffentlicher digital-kommunikativer Plattformen dringend geboten. Aber auch in anderen Sektoren wie der Verwaltung, der Bildung oder der Medizin vermindert eine zu große Abhängigkeit von wenigen Systemen oder Akteuren potenziell die individuelle wie kollektive Handlungsfähigkeit.

10.8 Querschnittsthema 8: Pfadabhängigkeiten, Zweitverwertung und Missbrauchsgefahren

Eine wichtige und oft vernachlässigte Rolle bei der Entwicklung und Nutzung von Technologien sind Pfadabhängigkeiten. Entscheidungen, die zu Beginn einer bestimmten

Entwicklung getroffen wurden, können noch lange nachwirken und sind teils schwer wieder aufzuheben, auch wenn sich der Kontext der Nutzung möglicherweise geändert hat. Dies gilt insbesondere dann, wenn es sich hierbei um grundlegende Technologien handelt, die Infrastrukturen prägen, welche in komplexe soziotechnische Zusammenhänge eingebettet sind, oder aber es durch die Nutzung bereits zu Gewöhnungseffekten gekommen ist. Von besonderer Relevanz für Infrastrukturen sind hier die Festlegungen von Normen und Standards, die zukünftige Entwicklungsmöglichkeiten begrenzen. Man denke hier etwa an standardisierte Stecker und Steckdosen, deren aktuelle Form nicht notwendigerweise besser oder schlechter als Alternativen sind, deren Änderung aber mit hohen Kosten verbunden wäre. Ein klassisches Beispiel für Pfadabhängigkeiten und Beharrungstendenzen durch Gewöhnungseffekte ist auch die Anordnung der Buchstaben auf Tastaturen. Die heute noch weit verbreitete QWERTY/QWERTZ-Tastaturbelegung wurde für mechanische Schreibmaschinen entwickelt, um die am häufigsten vorkommenden Buchstabenfolgen räumlich aufzuteilen, damit sich die mechanischen Typenhebel dieser Buchstaben weniger verhaken. Dass diese vergleichsweise wenig ergonomische Anordnung weiter Bestand hat, liegt an der Gewöhnung durch die Nutzenden.

Eine Konsequenz aus dieser Beobachtung lautet, dass grundlegenden Entscheidungen bei der Gestaltung von Infrastrukturen und dem Festsetzen von Standards und Normen eine hohe Aufmerksamkeit zukommen sollte. Für digitale Technologien, die zunehmend den Status kritischer Infrastrukturen annehmen, gilt dies umso mehr. Für Technologien, die entweder über einen weiten Verbreitungsgrad verfügen und/oder eine infrastrukturelle Dimension haben, könnte es zunehmend schwer und kostspielig werden, Änderungen einzufordern. Das Vorhandensein von Technologien und Infrastrukturen kann zudem selbst auch Erwartungen hinsichtlich der Verfügbarkeit und möglichst umfänglicher Nutzung wecken. Gerade bei kostspieligen Technologien dürfte eine

Tendenz auszumachen sein, deren Möglichkeiten voll auszuschöpfen – auch über das ursprüngliche Anwendungsfeld hinaus.

Eine „Verführung", technologische Möglichkeiten auch zu anderen Zwecken als zunächst intendiert zu nutzen, ist nicht prinzipiell problematisch. Mit dem Computer als Universalmaschine zeichnen sich digitale Technologien in der Regel durch vielfältige Einsatzszenarien aus. Dennoch öffnet dies auch die Tür für fremdnützigen Gebrauch oder Missbrauch. So wurden im Kontext der Coronapandemie weltweit sehr unterschiedliche Contact-Tracing-Apps entwickelt. Während in Deutschland auf ein dezentrales Modell gesetzt wurde, um möglichen Missbrauch bereits technisch zu verhindern, gab es aus anderen Ländern Hinweise auf Zweckentfremdung.[434]

Sobald eine Technologie etabliert ist, kann es schwer sein, weitere, auch missbräuchliche Nutzungsszenarien auszuschließen. Diese Tendenz wird oft auch mit dem Potenzial des sogenannten Dual Use beschrieben. Der Begriff wurde ursprünglich dafür verwendet anzuzeigen, dass Technologien sowohl für zivile als auch militärische Zwecke genutzt werden können. Mittlerweile wird der Begriff aber breiter verwendet, um darauf zu verweisen, dass Technologien oder Forschungsergebnisse sowohl für friedliche und nützliche Zwecke als auch zur absichtlichen Schädigung von Gesellschaft oder Umwelt, beispielsweise in krimineller oder terroristischer Absicht, eingesetzt werden können.[435] In den letzten Jahren sind digitale Technologien im Allgemeinen und KI im Besonderen zunehmend unter dem Blickwinkel von Dual Use beleuchtet worden. Dabei zeigt sich, dass die Konzeption von Dual Use hier um weitere Komplexitätsstufen ergänzt werden sollte, da digitale

434 So wurden Contact-Tracing-Apps, welche zur Nachverfolgung von COVID-19 entwickelt wurden, zum Beispiel auch zur Unterstützung polizeilicher Ermittlungen in Singapur herangezogen (https://www.theguardian.com/world/2021/jan/05/singapore-says-police-will-be-given-access-to-covid-19-contact-tracing-data [01.02.2023]).

435 Deutscher Ethikrat 2014; siehe auch https://www.sicherheitsrelevante-forschung.org [01.02.2023].

Technologien und insbesondere Grundlagentechnologien wie das maschinelle Lernen, oft sehr mannigfaltige Nutzungsmöglichkeiten eröffnen, in denen die Frage der Abgrenzung von Ge- und Missbrauch zunehmend schwieriger wird.

Empfehlung

» *Empfehlung Querschnittsthema 8:* Bei Technologien mit großen Auswirkungen oder hohem Verbreitungsgrad und vor allem dort, wo sich eine Nutzung von Technologien kaum oder gar nicht vermeiden lässt, müssen bereits zu Beginn der Entwicklungsplanung mögliche Langzeitfolgen wie Pfadabhängigkeiten im Allgemeinen sowie Dual-Use-Potenziale im Speziellen regelhaft und explizit mitgedacht und antizipiert werden. Dies gilt in besonderem Maße in der Anwendungsplanung. Dabei sind neben direkten, sektorspezifischen Schadenspotenzialen auch etwaige – natürlich deutlich schwieriger fass- und antizipierbare – sektorübergreifende Effekte zu bedenken. Hohe Standards für die Sicherheit und den Schutz der Privatsphäre (Security by Design, Privacy by Design) können ebenfalls dazu beitragen, spätere missbräuchliche Anwendungen einzuhegen bzw. möglichst zu verhindern. Bei besonders invasiven Technologien beispielsweise in der öffentlichen Verwaltung, die Bürgerinnen und Bürger gegebenenfalls verpflichtend nutzen müssen, sind besonders hohe Standards einzuhalten. Um dies sicherzustellen und überprüfen zu können, sind gegebenenfalls Open-Source-Ansätze angezeigt.

10.9 Querschnittsthema 9: Bias und Diskriminierung

Datenbasierte KI-Systeme lernen auf Basis vorhandener Daten. Resultierende Prognosen und Empfehlungen schreiben somit die Vergangenheit in die Zukunft fort, wodurch

Stereotype, aber auch bestehende gesellschaftliche Ungerechtigkeiten durch den Einbau in scheinbar neutrale Technologien reproduziert und sogar verstärkt werden können. In den letzten Jahren wurden die teils diskriminierenden Effekte insbesondere datenbasierter Technologien zur Entscheidungsunterstützung in zahlreichen Sektoren nachgewiesen. Für den Bereich der öffentlichen Verwaltung sei hier exemplarisch auf die Debatten rund um die Software COMPAS hingewiesen (vgl. Abschnitt 8.3.1). Auch im Kontext der für öffentliche Kommunikation und Meinungsbildung zentralen sozialen Medien und Suchmaschinen konnte gezeigt werden, dass algorithmische Systeme gesellschaftliche Stereotype und Ungerechtigkeiten reproduzieren können und dies in systematischen diskriminierenden Verzerrungen resultieren kann. In der Medizin wiederum gibt es zahlreiche Beispiele, dass verzerrte Trainingsdaten zu diskriminierenden Ergebnissen bei der Beurteilung von Patientinnen und Patienten durch KI-basierte Systeme führen können, so etwa bei der Einschätzung, wie viel Nachsorgebehandlungen Menschen nach einem Krankenhausaufenthalt benötigen – so wurde in diesem Fall die vorhandene Verzerrung in den Trainingsdaten in eine direkte Benachteiligung bestimmter Personengruppen übersetzt.[436] Auch im Bildungsbereich können systematische Verzerrungen nachgewiesen werden, insbesondere im Kontext von Audio- und Videoanalysen zum Zweck der Emotions- und Affekterkennung.

Die Ursachen für Diskriminierung durch KI-Systeme sind vielfältig. Oft liegt bei deren Entwicklung keine unmittelbare Diskriminierungsabsicht vor. Stattdessen sind diskriminierende Effekte das Resultat gesellschaftlicher Realitäten oder Stereotype in Kombination mit technisch-methodischen Entscheidungen wie beispielsweise der Wahl der Zielvariablen und Labels, der Auswahl der Trainingsdaten oder der verwendeten statistischen Analysemethoden. Dies erschwert

436 Obermeyer et al. 2019.

einerseits die Anwendung rechtlicher Regulierungen, die auf Vorsätzlichkeit von Diskriminierung bauen. Andererseits bedarf es detaillierter Analysen der methodisch-technischen Ursachen für diskriminierende Effekte, die häufig schwierig zu erkennen und nachzuweisen sind.

Von besonderer Bedeutung sind hier Trainingsdaten. Mangelnde Qualität, aber auch existierende gesellschaftliche Ungleichheiten, die sich in Daten widerspiegeln, können zu diskriminierenden Modellen führen. Über- und Unterrepräsentativität sind weitere Probleme, die diskriminierende Effekte haben können. Wenn Software im Bereich der Medizin zur Diagnostik von Hautkrebs vor allem auf Bildern mit heller Haut trainiert wurde, kann dies zu unterschiedlicher Genauigkeit bei der Befundung von verschiedenen Hautfarben führen. Die Folge wären beispielsweise überproportional häufigere Fehldiagnosen für Patientinnen und Patienten mit dunklerer Hautfarbe.

Das Gegenteil dieser mangelnden Berücksichtigung von Personengruppen in Datensätzen ist die Überrepräsentativität. Ein Beispiel ist der Einsatz von Software zur Vorhersage von Straftaten im Kontext von prädiktiver Polizeiarbeit. Falls eine Software eine bestimmte Gegend aufgrund bisheriger Straftaten als Hochrisikozone kategorisiert, werden dort gegebenenfalls in der Folge Polizeikontrollen verstärkt. Gerade aufgrund der erhöhten Kontrolle können dort dann noch mehr Straftaten verzeichnet werden. Im Gegenzug bleiben gegebenenfalls Straftaten in Niedrigrisikozonen aufgrund ausbleibender Kontrollen unerkannt. Durch beide Effekte kann sich die Datenlage weiter zuungunsten der Bewohnerschaft einer Hochrisikozone verschieben – nämlich für jene, die durch verstärkte Kontrollen unter Umständen ungerechtfertigt stigmatisiert werden.

Darüber hinaus besteht das Problem der sogenannten redundanten Encodierung: Sensible Attribute wie beispielsweise Geschlecht, religiöse Zugehörigkeit oder sexuelle Orientierung lassen sich teilweise aus anderen Datenpunkten, etwa

Bewegungsprofilen, Details des Medienkonsums sowie Angaben über Wohnort oder Hobbys, ableiten. Dadurch können diese Daten als sogenannte Proxy- bzw. Stellvertretervariablen für die geschützten Variablen fungieren. Das Resultat ist, dass Personen auch dann aufgrund ihres Geschlechts oder ihrer sexuellen oder religiösen Orientierung von der Software diskriminiert werden können, wenn diese Angaben gar nicht erhoben wurden, eben weil diese Kategorien aus anderen erhobenen Daten ableitbar sind.

In den zuvor genannten Beispielen ist Diskriminierung der Effekt von technisch-methodischen Entscheidungen, aber nicht notwendigerweise intendiert. Es ist allerdings zumindest denkbar, dass auch explizite Diskriminierungsabsichten in komplexen Systemen versteckt werden könnten. Dies gilt umso mehr für proprietäre, das heißt rechtlich geschützte Software, in welche die Personen, die sie verwenden, nicht nur aufgrund von technischer Komplexität, sondern auch aus rechtlichen Gründen keine Einsicht haben.

Empfehlung

» *Empfehlung Querschnittsthema 9:* Zum Schutz vor Diskriminierung in Anbetracht der zuvor dargelegten Herausforderungen bedarf es *angemessener Aufsicht und Kontrolle* von KI-Systemen. Besonders in sensiblen Bereichen erfordert dies den Auf- oder Ausbau gut ausgestatteter Institutionen. Hier gilt: je größer die Eingriffstiefe und je unumgänglicher die Systeme, desto höher die Anforderungen an Diskriminierungsminimierung. Auch bereits bei der Entwicklung von Technologien gilt es, Diskriminierung zu minimieren bzw. Fairness, Transparenz und Nachvollziehbarkeit herzustellen. Dies sollte sowohl durch Anreize – etwa Forschungsförderung – als auch durch entsprechende gesetzliche Anforderungen befördert werden, etwa hinsichtlich der Offenlegung, welche Maßnahmen zur Diskriminierungsminimierung bei der Softwareentwicklung ergriffen

wurden.[437] Allerdings haben technische wie regulatorische Maßnahmen zur Minimierung von Diskriminierung ihre Grenzen, unter anderem weil unterschiedliche Fairnessziele technisch nicht gleichzeitig erfüllt werden können. Es müssen also zugleich ethische und politische Entscheidungen getroffen werden, welche Kriterien für Gerechtigkeit in welchem Kontext zum Tragen kommen sollen. Diese Entscheidungen dürfen nicht den Personen, die Software entwickeln, und anderen direkt Beteiligten überlassen werden. Stattdessen bedarf es der Entwicklung geeigneter Verfahren und Institutionen, um diese Kriterien kontextspezifisch und demokratisch, gegebenenfalls immer wieder neu auszuhandeln. Je nach Anwendungskontext und Sensibilität des einzusetzenden Systems kann die Beteiligung der Öffentlichkeit erforderlich sein. Dabei sollte der Schutz der jeweils bedürftigsten bzw. von Entscheidungen besonders betroffenen Gruppen besonders berücksichtigt werden.

10.10 Querschnittsthema 10: Transparenz und Nachvollziehbarkeit – Kontrolle und Verantwortung

KI-Systeme sind mitunter wenig transparent und nachvollziehbar. Diese Opazität hat verschiedene Ursachen, die vom Schutz geistigen Eigentums über die Komplexität und Nichtnachvollziehbarkeit der Verfahren bis hin zur mangelnden Durchsichtigkeit von Entscheidungsstrukturen, in die der Einsatz algorithmischer Systeme eingebettet ist, reichen. Als Reaktion auf diese vielfältigen Herausforderungen gibt es Bemühungen, die Transparenz und Nachvollziehbarkeit

437 Der derzeit diskutierte Vorschlag der Europäischen Kommission für einen Artificial Intelligence Act verfolgt bereits diesen Ansatz, auf der einen Seite die Forschung zu KI-Technologien zu fördern und auf der anderen, einen rechtlichen Rahmen für ihre Entwicklung und Anwendung zu schaffen.

durch technische, organisatorische und rechtliche Mittel zu erhöhen.[438]

Fragen von Transparenz, Erklärbarkeit und Nachvollziehbarkeit sind mit Fragen von Kontrolle und Verantwortung verbunden. Zwar besteht zwischen ihnen ein gewisser Zusammenhang, doch sind Transparenz und Nachvollziehbarkeit algorithmischer Systeme für die Kontrolle und die Verantwortung für ihren Einsatz weder zwingend notwendig noch hinreichend.

Einerseits kann es auch bei prinzipiell transparenten und nachvollziehbaren Methoden wie zum Beispiel Entscheidungsbäumen dazu kommen, dass verantwortliches Handeln und angemessene Kontrolle ausbleiben. Zudem ist in Bezug auf Offenlegungspraktiken auf das Problem strategischer Transparenz hinzuweisen. So könnten insbesondere im Bereich der öffentlichen Kommunikation und Meinungsbildung Plattformbetreiber entweder irrelevante und unzureichende Informationen transparent machen (z. B.in Bezug auf Prozesse und Effekte von Content-Moderation) oder aber relevante Informationen unter einer Fülle irrelevanter Information verbergen. In diesen Fällen würde Transparenz also nicht zwingend zu verantwortlichem Handeln und Kontrolle führen.

Andererseits sind Kontrolle und Verantwortung auch ohne vollständige Transparenz möglich. So kann bei der Nutzung von Softwaresystemen, die auf Deep-Learning-Ansätzen beruhen, Herstellern oder Anwendern die volle Verantwortung für den Einsatz dieser Systeme zugewiesen werden, selbst wenn ihnen die Details der Verarbeitung unbekannt sind. Sie trügen dann die Verantwortung, derartige Systeme zum Einsatz gebracht zu haben, und müssten begründen, warum diese Intransparenz akzeptabel ist – etwa weil der mögliche Schaden gering oder der zusätzliche Nutzen dieser Systeme (z. B. in

438 Schlagworte dieser Debatte sind neben Transparenz (*transparency*) insbesondere auch Erklärbarkeit (*explainability/explainable AI/explicability*), Beobachtbarkeit (*observability*) und Nachvollziehbarkeit sowie Verantwortung und Haftung (*responsibility*, *accountability* und *liability*).

Bezug auf eine höhere Genauigkeit der Prognosen) die Nachteile der Intransparenz überwiegt. So kann es einerseits sein, dass im medizinischen Kontext aus guten Gründen Software in der Krebsdiagnostik eingesetzt wird, deren Prognosen durch diejenigen, die sie einsetzen, zwar nicht mehr nachvollziehbar erklärt werden können, deren höhere Genauigkeit jedoch diesen Nachteil überwiegt. Umgekehrt könnte in anderen Kontexten bzw. bei sensiblen Entscheidungen in der öffentlichen Verwaltung die Notwendigkeit der vollständig nachvollziehbaren Begründung von Entscheidungen dazu führen, dass intransparente Systeme nicht eingesetzt werden dürfen.

Anforderungen an Transparenz, Erklärbarkeit und Nachvollziehbarkeit sind dementsprechend in Abhängigkeit von den jeweiligen Zielen, die mit Transparenz und Nachvollziehbarkeit verfolgt werden, nach den Personen, die Informationen erhalten, und nach dem jeweiligen Anwendungskontext zu konkretisieren. Eine Nutzerin, die Auskunft darüber verlangt, warum sie keinen Kredit bekommen hat, bedarf einer anderen Art der Erklärung als eine Aufsichtsbehörde, die prüfen muss, ob eine Software in ihren Prognosen der Kreditwürdigkeit systematisch Frauen diskriminiert. Besonders hohe Anforderungen gelten für Systeme, die in hochsensiblen Bereichen eingesetzt werden, beispielsweise bei Entscheidungen mit hoher Tragweite für das Leben von Menschen. Auch dort, wo Systeme eine Quasi-Monopolstellung erlangen, sind hohe Anforderungen an Transparenz, Erklärbarkeit und Nachvollziehbarkeit zu stellen, die möglicherweise den Einsatz von besonders intransparenten Systemen (z. B. basierend auf Deep Learning) ausschließen und nach nachvollziehbareren Verfahren verlangen (z. B. Entscheidungsbäume). Bei der Entwicklung der Software wiederum müssen technische und organisatorische Voraussetzungen geschaffen und eingefordert werden, beispielsweise durch Dokumentations- und Offenlegungspflichten, damit diese Erklärungen später überhaupt erst geliefert werden können.

Die Spezifizierung und ausgewogene Ausgestaltung von Offenlegungspflichten stellen eine besondere Herausforderung dar, muss doch sichergestellt werden, dass einerseits relevante Informationen geteilt werden, aber andererseits weder die Geschäftsinteressen von Anbietern über Gebühr unterminiert noch Flanken für Angriffe und Sicherheitslücken eröffnet werden.

Empfehlung

» *Empfehlung Querschnittsthema 10:* Es bedarf der Entwicklung ausgewogener aufgaben-, adressaten- und kontextspezifischer Standards für Transparenz, Erklärbarkeit und Nachvollziehbarkeit und ihrer Bedeutung für Kontrolle und Verantwortung sowie für deren Umsetzung durch verbindliche technische und organisatorische Vorgaben. Dabei muss den Anforderungen an Sicherheit und Schutz vor Missbrauch, Datenschutz sowie dem Schutz von intellektuellem Eigentum und Geschäftsgeheimnissen in angemessener Weise Rechnung getragen werden. Je nach Kontext sind hier unterschiedliche Zeitpunkte (ex ante, ex post, Realtime) sowie unterschiedliche Verfahren und Grade der Offenlegung zu spezifizieren.

10.11 Fazit

Im Rahmen dieser Stellungnahme wurden die Auswirkungen einer zunehmenden Delegation menschlicher Tätigkeiten an digitale Technologien, insbesondere KI-basierte Softwaresysteme, analysiert. In zahlreichen Beispielen aus den Bereichen der Medizin, der schulischen Bildung, der öffentlichen Kommunikation und Meinungsbildung sowie der öffentlichen Verwaltung zeigte sich, dass dieses Delegieren sowohl mit Erweiterungen als auch mit Verminderungen menschlicher Handlungsmöglichkeiten einhergeht und sich dadurch sowohl förderlich als auch hinderlich auf die Realisierung menschlicher

Autorschaft auswirken kann. Eine Berücksichtigung dieser Auswirkungen sollte daher jedweder Entscheidung über eine Delegation menschlicher Tätigkeiten an Softwaresysteme – bis möglicherweise hin zu einer vollständigen Ersetzung des Menschen durch algorithmische Systeme – vorausgehen.

Ziel und Richtschnur ethischer Bewertung muss hierbei immer die Erhöhung menschlicher Autorschaft sein. Dabei ist zu berücksichtigen, dass die Erweiterung von Handlungsmöglichkeiten für eine Personengruppe mit deren Verminderung für andere einhergehen kann. Diesen Unterschieden ist Rechnung zu tragen, insbesondere im Hinblick auf den Schutz und die Verbesserung der Lebensbedingungen vulnerabler oder benachteiligter Gruppen. Die hier vorgelegte Analyse hat zahlreiche übergreifende Themen offengelegt, die in allen vier untersuchten Sektoren aufscheinen – wenn auch nicht immer in gleicher Art und Weise. Letztlich zeigt sich, dass die normativen Ansprüche an die Gestaltung und den Einsatz solcher Technologien, beispielsweise in Bezug auf Anforderungen hinsichtlich Transparenz und Nachvollziehbarkeit, den Schutz der Privatsphäre sowie die Verhinderung von Diskriminierung, zwar in allen Bereichen und für alle Betroffenen von hoher Bedeutung sind, sie jedoch sektor-, kontext- und adressatenspezifisch konkretisiert werden müssen, um angemessen zu sein und wirksam werden zu können.

LITERATURVERZEICHNIS

Ad Hoc Committee on Artificial Intelligence (2021): Possible Elements of a Legal Framework on Artificial Intelligence, Based on the Council of Europe's Standards on Human Rights, Democracy and the Rule of Law. https://rm.coe.int/cahai-2021-09rev-elements/1680a6d90d [19.01.2023].

Ahmed, H. U. et al. (2017): Diagnostic accuracy of multi-parametric MRI and TRUS biopsy in prostate cancer (PROMIS): a paired validating confirmatory study. In: The Lancet, 389 (10071), 815–822 (DOI: 10.1016/S0140-6736(16)32401-1).

AlgorithmWatch (2020): Automating Society Report 2020. https://automatingsociety.algorithmwatch.org/wp-content/uploads/2020/12/Automating-Society-Report-2020.pdf [22.02.2023].

AlgorithmWatch (2019): Automating Society. Taking Stock of Automated Decision-Making in the EU. https://algorithmwatch.org/en/wp-content/uploads/2019/02/Automating_Society_Report_2019.pdf [22.02.2023].

Allem, J.-P. (2020): Social media fuels wave of coronavirus misinformation as users focus on popularity, not accuracy. In: The Conversation, 06.04.2020. https://theconversation.com/social-media-fuels-wave-of-coronavirus-misinformation-as-users-focus-on-popularity-not-accuracy-135179 [12.01.2023].

Allhutter, D. et al. (2020a): Der AMS-Algorithmus. Eine Soziotechnische Analyse des Arbeitsmarktchancen-Assistenz-Systems (AMAS). https://epub.oeaw.ac.at/ita/ita-projektberichte/2020-02.pdf [08.02.2023].

Allhutter, D. et al. (2020b): Algorithmic profiling of job seekers in Austria: how austerity politics are made effective. In: Frontiers in Big Data, 3, Art.-Nr. 5 (DOI: 10.3389/fdata.2020.00005).

Alston, J. (2019): Report of the Special Rapporteur on Extreme Poverty and Human Rights. UN Doc. A/74/493. https://undocs.org/A/74/493 [22.02.2023].

Alzoubi, D. et al. (2021): TeachActive feedback dashboard: using automated classroom analytics to visualize pedagogical strategies at a glance. In: CHI'21 Extended Abstracts, Art.-Nr. 312 (DOI: 10.1145/3411763.3451709).

Andrews, D. A.; Bonta, J. (2010): The Psychology of Criminal Conduct. Abingdon.

Aneesh, A. (2009): Global labor: algocratic modes of organization. In: Sociological Theory, 27 (4), 347–370 (DOI: 10.1111/j.1467-9558.2009.01352.x).

Arnold, C.; Günther, J. (Hg.) (2022): Arbeitsrecht 4.0. Praxishandbuch zum Arbeits-, IP- und Datenschutzrecht in einer digitalisierten Arbeitswelt (2. Auflage). München.

Bainbridge, L. (1983): Ironies of automation. In: Automatica, 19 (6), 775–779 (DOI: 10.1016/0005-1098(83)90046-8).

Bakshy, E.; Messing, S.; Adamic, L. A. (2015): Exposure to ideologically diverse news and opinion on Facebook. In: Science, 348 (6239), 1130–1132 (DOI: 10.1126/science.aaa1160).

Baltes, P. B. (1987): Theoretical propositions of life-span developmental psychology: on the dynamics between growth and decline. In: Developmental Psychology, 23 (5), 611–626 (10.1037/0012-1649.23.5.611).

Baltzer, P. A. T. (2021): Künstliche Intelligenz in der Mammadiagnostik. Anwendungsgebiete aus klinischer Perspektive. In: Der Radiologe, 61 (2), 192–198 (DOI: 10.1007/s00117-020-00802-2).

Barocas, S.; Selbst, A. D. (2016): Big data's disparate impact. In: California Law Review, 104 (3), 671–732 (DOI: 10.15779/Z38BG31).

Bastian, P. (2012): Die Überlegenheit statistischer Urteilsbildung im Kinderschutz. Plädoyer für einen Perspektivwechsel hin zu einer angemessenen Form sozialpädagogischer Diagnosen. In: Marthaler, T. et al. (Hg.): Rationalitäten des Kinderschutzes. Kindeswohl und soziale Interventionen aus pluraler Perspektive. Wiesbaden, 249–267.

Bastian, P.; Freres, K.; Schrödter, M. (2017): Risiko und Sicherheit als Orientierung im Kinderschutz. Deutschland und USA im Vergleich. In: Soziale Passagen, 9 (2), 245–261.

Bastian, P. et al. (2018): Bauchgefühle in der Sozialen Arbeit. In: Kommission Sozialpädagogik (Hg.): Wa(h)re Gefühle? Sozialpädagogische Emotionsarbeit im wohlfahrtsstaatlichen Kontext. Weinheim, 128–140.

Beauchamp, T. L.; Childress, J. F. (2001): Principles of Biomedical Ethics. New York (NY).

Beck, I.; Greving, H. (2012): Lebenswelt, Lebenslage. In: dies. (Hg.): Lebenslage und Lebensbewältigung. Stuttgart, 15–59.

Belpaeme, T.; Tanaka, F. (2021): Social robots as educators. In: OECD Digital Education Outlook 2021. Pushing the Frontiers with Artificial Intelligence, Blockchain and Robots. Paris, 143–158.

Bennett, M. et al. (2007): Neuroscience and Philosophy. Brain, Mind, and Language. New York (NY).

Bickle, J. (2020): Multiple realizability. In: Stanford Encyclopedia of Philosophy (Summer 2020 Edition). https://plato.stanford.edu/archives/sum2020/entries/multiple-realizability [04.01.2023].

Bijker, W. E.; Hughes, T. P.; Pinch, T. J. (Hg.) (1987): The Social Construction of Technological Systems. New Directions in the Sociology and History of Technology. Cambridge (MA).

Binet, A.; Simon, T. (1905): Méthodes nouvelles pour le diagnostic du niveau intellectuel des anormaux. In: L'Année Psychologique, 11, 191–244 (DOI: 10.3406/psy.1904.3675).

Bitkom (2017): Künstliche Intelligenz. Wirtschaftliche Bedeutung, gesellschaftliche Herausforderungen, menschliche Verantwortung. https://www.bitkom.org/sites/main/files/file/import/171012-KI-Gipfelpapier-online.pdf [07.02.2023].

Block, N. (1978): Troubles with functionalism. In: Savage, C. W. (Hg.): Perception and Cognition. Issues in the Foundations of Psychology. Minneapolis (MN), 261–325.

Bodó, B.; Helberger, N.; de Vreese, C. H. (2017): Political micro-targeting: a Manchurian candidate or just a dark horse? In: Internet Policy Review, 6 (4) (DOI: 10.14763/2017.4.776).

Bomhard, D.; Merkle, M. (2021): Regulation of artificial intelligence. The EU Commission's proposal of an AI Act. In: Journal of European Consumer and Market Law, 10 (6), 257–261.

Bormann, F.-J. (2021): Ist die praktische Vernunft des Menschen durch KI-Systeme ersetzbar? Zum unterschiedlichen Status von menschlichen Personen und (selbst-)lernenden Maschinen. In: Fritz, A. et al. (Hg.) (2021): Digitalisierung im Gesundheitswesen. Anthropologische und ethische Herausforderungen der Mensch-Maschine-Interaktion. Freiburg im Breisgau, 41–64.

Bottek, C. (2014): Unterlassungen und ihre Folgen. Handlungs- und kausalitätstheoretische Überlegungen. Tübingen.

Bovens, L.; Hartmann, S. (2003): Bayesian Epistemology. Oxford.

Bowker, G. C.; Star, S. L. (1999): Sorting Things Out. Classification and Its Consequences. Cambridge (MA).

Bradshaw, S.; DiResta, R.; Miller, C. (2022): Playing both sides: Russian state-backed media coverage of the #BlackLivesMatter movement. In: The International Journal of Press/Politics (DOI: 10.1177/19401612221082052).

Bradshaw, S.; Henle, A. (2021): The gender dimensions of foreign influence operations. In: International Journal of Communication, 15, 4596–4618.

Brady, W. J. et al. (2017): Emotion shapes the diffusion of moralized content in social networks. In: PNAS, 114 (28), 7313–7318 (DOI: 10.1073/pnas.1618923114).

Brown, T. B. et al. (2020): Language Models are Few-Shot Learners. https://arxiv.org/abs/2005.14165 [10.02.2023].

Büchi, M.; Festic, N.; Latzer, M. (2022): The chilling effects of digital dataveillance: a theoretical model and an empirical research agenda. In: Big Data & Society, 9 (1) (DOI: 10.1177/20539517211065368).

Buckley, R. P. et al. (2021): Regulating artificial intelligence in finance: putting the human in the loop. In: Sydney Law Review, 43 (1), 43–81.

Bundesministerium für Klimaschutz, Umwelt, Energie, Mobilität, Innovation und Technologie (2021): Strategie der Bundesregierung für Künstliche Intelligenz. Artificial Intelligence Mission Austria 2030 (AIM AT 2030). Wien.

Burchardi, S. (2022): Risikotragung für KI-Systeme. In: Europäische Zeitschrift für Wirtschaftsrecht, 33 (15), 685–692.

Burhan, R.; Moradzadeh, J. (2020): Neurotransmitter dopamine (DA) and its role in the development of social media addiction. In: Journal of Neurology & Neurophysiology, 11 (7), Art.-Nr. 507. https://www.iomcworld.org/open-access/neurotransmitter-dopamine-da-and-its-role-in-the-development-of-social-media-addiction-59222.html [13.01.2023].

Butz, F. et al. (2021): Automatisierte Risikoprognosen im Kontext von Bewährungsentscheidungen. In: Bewährungshilfe, 68 (3), 241–259.

Bynum, T. (2015): Computer and information ethics. In: Stanford Encyclopedia of Philosophy (Summer 2018 Edition). https://plato.stanford.edu/archives/sum2018/entries/ethics-computer [20.12.2022].

Carbonell, J. G.; Michalski, R. S.; Mitchell, T. M. (1983): Machine learning: a historical and methodological analysis. In: AI Magazine, 4 (3), 69–78 (DOI: 10.1609/aimag.v4i3.406).

Castanho Silva, B.; Vegetti, F; Littvay, L. (2017): The elite is up to something: exploring the relation between populism and belief in conspiracy theories. In: Swiss Political Science Review, 23 (4), 423–443 (DOI: 10.1111/spsr.12270).

Cheng, C. et al. (2021): Prevalence of social media addiction across 32 nations: meta-analysis with subgroup analysis of classification schemes and cultural values. In: Addictive Behaviors, 117, Art.-Nr. 106845 (DOI: 10.1016/j.addbeh.2021.106845).

Chuai, Y.; Zhao, J. (2022): Anger can make fake news viral online. In: Frontiers in Physics, 10, Art.-Nr. 970174 (DOI: 10.3389/fphy.2022.970174).

Cinelli, M. et al. (2021): The echo chamber effect on social media. In: PNAS, 118 (9), Art.-Nr. e2023301118 (DOI: 10.1073/pnas.2023301118).

Clark, A. (2012): Embodied, embedded, and extended cognition. In: Frankish, K.; Ramsey, W. M. (Hg.): The Cambridge Handbook of Cognitive Science. Cambridge, 275–291.

Cornel, H.; Pruin, I. (2021): Die Implementierung der Risikoorientierung in den Bundesländern. In: Cornel, H.; Kawamura-Reindl, G. (Hg.): Bewährungshilfe. Theorie und Praxis eines Handlungsfeldes Sozialer Arbeit. Weinheim, 105–118.

Cremers, A. B. et al. (2019): Vertrauenswürdiger Einsatz von Künstlicher Intelligenz. Handlungsfelder aus philosophischer, ethischer, rechtlicher und technologischer Sicht als Grundlage für eine Zertifizierung von Künstlicher Intelligenz. https://www.iais.fraunhofer.de/content/dam/iais/KINRW/Whitepaper_KI-Zertifizierung.pdf [09.01.2023].

Croes, E. A. J.; Antheunis, M. L. (2021): 36 questions to loving a chatbot: are people willing to self-disclose to a chatbot? In: Følstad, A. et al. (Hg.): Chatbot Research and Design. 4th International Workshop, Conversations 2020. Cham, 81–95.

Danaher, J. (2016): The threat of algocracy: reality, resistance and accommodation. In: Philosophy & Technology, 29 (3), 245–268.

Datenethikkommission der Bundesregierung (2019): Gutachten der Datenethikkommission. Berlin.

Davidson, T.; Bhattacharya, D.; Weber, I. (2019): Racial bias in hate speech and abusive language detection datasets. In: Proceedings of the Third Workshop on Abusive Language Online, 25–35 (DOI: 10.18653/v1/W19-3504).

Deary, I. J. (2020): Intelligence. A Very Short Introduction. Oxford.

Dechsling, R. (1989): Das Verhältnismässigkeitsgebot. Eine Bestandsaufnahme der Literatur zur Verhältnismässigkeit staatlichen Handelns. München.

Del Vicario, M. et al. (2016): The spreading of misinformation online. In: PNAS, 113 (3), 554–559 (DOI: 10.1073/pnas.1517441113).

Denga, M. (2018): Deliktische Haftung für künstliche Intelligenz. Warum die Verschuldenshaftung des BGB auch künftig die bessere Schadensausgleichsordnung bedeutet. In: Computer und Recht, 34 (2), 69–78 (DOI: 10.9785/cr-2018-0203).

Dennis, A. R. et al. (2020): User reactions to COVID-19 screening chatbots from reputable providers. In: Journal of the American Medical Informatics Association, 27 (11), 1727–1731 (DOI: 10.1093/jamia/ocaa167).

Deutscher Bundestag (2020): Bericht der Enquete-Kommission Künstliche Intelligenz – Gesellschaftliche Verantwortung und wirtschaftliche, soziale und ökologische Potenziale. Drucksache 19/23700. https://dserver.bundestag.de/btd/19/237/1923700.pdf [18.01.2023].

Deutscher Bundestag (2017): Wortprotokoll der 114. Sitzung des Innenausschusses. Protokoll-Nr. 18/114. https://www.bundestag.de/resource/blob/515144/87bdd46c572ebeb4863d139812f3fc4a/Protokoll-114-data.pdf [03.03.2023].

Deutscher Ethikrat (2022a): Pandemie und psychische Gesundheit. Aufmerksamkeit, Beistand und Unterstützung für Kinder, Jugendliche und junge Erwachsene in und nach gesellschaftlichen Krisen. Berlin.

Deutscher Ethikrat (2022b): Vulnerabilität und Resilienz in der Krise – Ethische Kriterien für Entscheidungen in einer Pandemie. Berlin.

Deutscher Ethikrat (2020a): Robotik für gute Pflege. Berlin.

Deutscher Ethikrat (2020b): Tierwohlachtung – Zum verantwortlichen Umgang mit Nutztieren. Berlin.

Deutscher Ethikrat (2017): Big Data und Gesundheit – Datensouveränität als informationelle Freiheitsgestaltung. Berlin.

Deutscher Ethikrat (2014): Biosicherheit – Freiheit und Verantwortung in der Wissenschaft. Berlin.

Dierkes, M.; Hoffmann, U.; Marz, L. (1992): Leitbild und Technik. Zur Entstehung und Steuerung technischer Innovationen. Berlin.

Dillenbourg, P. (2021): Classroom analytics: zooming out from a pupil to a classroom. In: OECD Digital Education Outlook 2021. Pushing the Frontiers with Artificial Intelligence, Blockchain and Robots. Paris, 105–122.

DiResta, R. et al. (2019): The Tactics & Tropes of the Internet Research Agency. https://www.intelligence.senate.gov/sites/default/files/documents/NewKnowledge-Disinformation-Report-Whitepaper.pdf [12.01.2023].

Dolata, U.; Werle, R. (Hg.) (2007): Gesellschaft und die Macht der Technik. Sozioökonomischer und institutioneller Wandel durch Technisierung. Frankfurt am Main.

Donelan, H. (2016): Social media for professional development and networking opportunities in academia. In: Journal of Further and Higher Education, 40 (5), 706–729 (DOI: 10.1080/0309877X.2015.1014321).

Drake, B. et al. (2020): A Practical Framework for Considering the Use of Predictive Risk Modeling in Child Welfare. In: The Annals of the American Academy of Political and Social Science, 692 (1), 162–181 (DOI: 10.1177/0002716220978200).

Dreyfus, H. L. (1992): What Computers Still Can't Do. A Critique of Artificial Reason. Cambridge (MA).

Dubois, E.; Blank, G. (2018): The echo chamber is overstated: The moderating effect of political interest and diverse media. In: Information, Communication & Society, 21 (5), 729–745 (DOI: 10.1080/1369118X.2018.1428656).

E-READ (2019): Stavanger Declaration Concerning the Future of Reading. https://ereadcost.eu/wp-content/uploads/2019/01/StavangerDeclaration.pdf [12.01.2023].

Ebers, M. et al. (Hg.) (2020): Künstliche Intelligenz und Robotik. Rechtshandbuch (1. Auflage). München.

Ebert, D. D. et al. (2018): Internet- and mobile-based psychological interventions: applications, efficacy, and potential for improving mental health. In: European Psychologist, 23 (2), 167–187 (DOI: 10.1027/1016-9040/a000318).

Egbert, S. (2020): Predictive Policing als Treiber rechtlicher Innovation? In: Zeitschrift für Rechtssoziologie, 40 (1–2), 26–51 (DOI: 10.1515/zfrs-2020-0002).

Egbert, S. (2018): Predictive Policing in Deutschland. Grundlagen, Risiken, (mögliche) Zukunft. In: Strafverteidigervereinigungen (Hg.): Räume der Unfreiheit. Berlin, 241–265.

Egbert, S.; Leese, M. (2020): Criminal Futures. Predictive Policing and Everyday Police Work. London.

Eifert, M. et al. (2021): Taming the giants: the DMA/DSA package. In: Common Market Law Review, 58 (4), 987–1028 (DOI: 10.54648/cola2021065).

Eisele, J.; Böhm, K. (2020): Potential und Risiken von Predictive Policing. In: Beck, S.; Kusche, C.; Valerius, B. (Hg.): Digitalisierung, Automatisierung, KI und Recht. Festgabe zum 10-jährigen Bestehen der Forschungsstelle RobotRecht. Baden-Baden, 519–533.

Ender, A. M. et al. (2023): The relationship between social media use and beliefs in conspiracy theories and misinformation. In: Political Behavior, 45 (2), 781–804 (DOI: 10.1007/s11109-021-09734-6).

Epping, V.; Hillgruber, C. (Hg.) (2022): BeckOK Grundgesetz (52. Edition, Stand: 15.08.2022). https://beck-online.beck.de/Dokument?vpath=bibdata/komm/beckokgg_52/cont/beckokgg.htm [17.01.2023].

Epstein, R.; Robertson, R. E. (2015): The search engine manipulation effect (SEME) and its possible impact on the outcomes of elections. In: PNAS, 112 (33), E4512–E4521 (DOI: 10.1073/pnas.1419828112).

Ethik-Kommission Automatisiertes und vernetztes Fahren (2017): Bericht. https://bmdv.bund.de/goto?id=348344 [18.01.2023].

Eubanks, V. (2018): Automating Inequality. How High-Tech Tools Profile, Police, and Punish the Poor. New York (NY).

Europäische Kommission (2020): Weißbuch zur Künstlichen Intelligenz – Ein europäisches Konzept für Exzellenz und Vertrauen. https://eur-lex.europa.eu/legal-content/DE/TXT/?uri=CELEX:52020DC0065 [09.01.2023].

Faber, J. M.; Luyten, H.; Visscher, A. J. (2017): The effects of a digital formative assessment tool on mathematics achievement and student motivation: results of a randomized experiment. In: Computers & Education, 106, 83–96 (DOI: 10.1016/j.compedu.2016.12.001).

Fan, R. et al. (2014): Anger is more influential than joy: sentiment correlation in Weibo. In: PLOS One, 9 (10), Art.-Nr. e110184 (DOI: 10.1371/journal.pone.0110184).

Ferrara, E. et al. (2020): Characterizing social media manipulation in the 2020 U.S. presidential election. In: First Monday, 25 (11) (DOI: 10.5210/fm.v25i11.11431).

Ferrara, E. et al. (2016): The rise of social bots. In: Communications of the ACM, 59 (7), 96–104 (DOI: 10.1145/2818717).

Fichtner, L. (2022): Content moderation and the quest for democratic legitimacy. In: Weizenbaum Journal of the Digital Society, 2 (2), Art.-Nr. w2.2.2 (DOI: 10.34669/wi.wjds/2.2.2).

Fiske, A.; Henningsen, P.; Buyx, A. (2019): Your robot therapist will see you now: ethical implications of embodied artificial intelligence in psychiatry, psychology, and psychotherapy. In: Journal of Medical Internet Research, 21 (5), Art.-Nr. e13216 (DOI: 10.2196/13216).

Fjeld, J. et al. (2020): Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-based Approaches to Principles for AI. https://ssrn.com/abstract=3518482 [10.02.2023].

Floridi, L. (2016): Faultless responsibility: on the nature and allocation of moral responsibility for distributed moral actions. In: Philosophical Transactions A, 374 (2083), Art.-Nr. 20160112 (DOI: 10.1098/rsta.2016.0112).

Floridi, L. (2013): Distributed morality in an information society. In: Science and Engineering Ethics, 19 (3), 727–743 (DOI: 10.1007/s11948-012-9413-4).

Floridi, L.; Sanders, J. W. (2004): On the morality of artificial agents. In: Minds and Machines, 14 (3), 349–379 (DOI: 10.1023/B:MIND.0000035461.63578.9d).

Floridi, L. et al. (2018): AI4People – an ethical framework for a good AI society: opportunities, risks, principles, and recommendations. In: Minds and Machines, 28 (4), 689–707 (DOI: 10.1007/s11023-018-9482-5).

Fodor, J. A. (1974): Special sciences (or: the disunity of science as a working hypothesis). In: Synthese, 28 (2), 97–115 (DOI: 10.1007/BF00485230).

Frankfurt, H. G. (1971): Freedom of the will and the concept of a person. In: The Journal of Philosophy, 68 (1), 5–20 (DOI: 10.2307/2024717).

Freund, G. (1992): Erfolgsdelikt und Unterlassen. Zu den Legitimationsbedingungen von Schuldspruch und Strafe. Köln.

Friedman, B.; Hendry, D. G. (2019): Value Sensitive Design. Shaping Technology with Moral Imagination. Cambridge (MA).

Friedman, B.; Nissenbaum, H. (1996): Bias in computer systems. In: ACM Transactions on Information Systems, 14 (3), 330–347 (DOI: 10.1145/230538.230561).

Frisch, W. (2016): Strafrecht und Solidarität – Zugleich zu Notstand und unterlassener Hilfeleistung. In: Goltdammer's Archiv für Strafrecht, 163 (3), 121–137.

Fuchs, T. (2013): Verkörperung, Sozialität und Kultur. In: Breyer, T. et al. (Hg.): Interdisziplinäre Anthropologie. Leib – Geist – Kultur. Heidelberg, 11–33.

Fuhrmann, M. et al. (1989): Person. In: Ritter, J.; Gründer, K. (Hg.): Historisches Wörterbuch der Philosophie. Band 7: P-Q. Basel, Sp. 269–338.

Garside, J.; Plana, L. A. (2020): The SpiNNaker chip. In: Furber, S.; Bogdan, P. (Hg.): SpiNNaker. A Spiking Neural Network Architecture. Boston (MA), 17–51.

Geminn, C. (2021): Die Regulierung Künstlicher Intelligenz. Anmerkungen zum Entwurf eines Artificial Intelligence Act. In: Zeitschrift für Datenschutz, 11 (7), 354–359.

Gerhardinger, S. (2020). Entwicklung der Therapeutenpersönlichkeit. Orientierungshilfen für Psychotherapeutinnen und Psychotherapeuten. Berlin.

Ghanem, C. (2021): Bewährungshilfe zwischen Risiko- und Ressourcenorientierung. In: Cornel, H.; Kawamura-Reindl, G. (Hg.): Bewährungshilfe. Theorie und Praxis eines Handlungsfeldes Sozialer Arbeit. Weinheim, 84–91.

Gibbard, A. (1973): Manipulation of voting schemes: a general result. In: Econometrica, 41 (4), 587–601 (DOI: 10.2307/1914083).

Gielen, N.; Uphues, S. (2021): Digital Markets Act und Digital Services Act. Regulierung von Markt- und Meinungsmacht durch die Europäische Union. In: Europäische Zeitschrift für Wirtschaftsrecht, 32 (14), 627–637.

Gigerenzer, G.; Mata, J.; Frank, R. (2009): Public knowledge of benefits of breast and prostate cancer screening in Europe. In: Journal of the National Cancer Institute, 101 (17), 1216–1220 (DOI: 10.1093/jnci/djp237).

Gillingham, P. (2021): Big Data, prädiktive Analytik und Soziale Arbeit. Ein Überblick. In: Sozial Extra, 45 (1), 31–35 (DOI: 10.1007/s12054-020-00348-6).

Gillingham, P. (2016): Predictive risk modelling to prevent child maltreatment and other adverse outcomes for service users: inside the 'black box' of machine learning. In: The British Journal of Social Work, 46 (4), 1044–1058 (DOI: 10.1093/bjsw/bcv031).

Gluba, A. (2017): Der Modus Operandi bei Fällen der Near Repeat-Victimisation. In: Kriminalistik, 71 (6), 369–375.

Good, J. (2021): Serving students with special needs better: how digital technology can help. In: OECD Digital Education Outlook 2021. Pushing the Frontiers with Artificial Intelligence, Blockchain and Robots. Paris, 123–142.

Görder, B. (2021): Die Macht der Muster. Die Ethik der Sozialen Arbeit vor professionsbezogenen und gesellschaftlichen Herausforderungen durch ‚künstliche Intelligenz'. In: Ethik Journal, 7 (2). https://www.ethikjournal.de/fileadmin/user_upload/ethikjournal/Texte_Ausgabe_2021_2/Goerder_Ethikjournal_2.2021.pdf [08.02.2023].

Görz, G.; Schmid, U.; Braun, T. (Hg.) (2021): Handbuch der Künstlichen Intelligenz (6. Auflage). Berlin.

Graesser, A. C. et al. (2012): AutoTutor. In: McCarthy, P. M.; Boonthum-Denecke, C. (Hg.): Applied Natural Language Processing: Identification, Investigation and Resolution. Hershey (PA), 169–187.

Gratzer, D.; Goldbloom, D. (2020): Therapy and e-therapy – preparing future psychiatrists in the era of apps and chatbots. In: Academic Psychiatry, 44 (2), 231–234 (DOI: 10.1007/s40596-019-01170-3).

Gray, C. C.; Perkins, D. (2019): Utilizing early engagement and machine learning to predict student outcomes. In: Computers & Education, 131, 22–32 (DOI: 10.1016/j.compedu.2018.12.006).

Grunwald, A. (2022): Technikfolgenabschätzung. Einführung. Baden-Baden.

Grunwald, A. (2021): Der homo responsibilis. Ein nachdenklicher Gang durch den Garten aktueller Erzählungen. In: ders. (Hg.): Wer bist du, Mensch? Transformationen menschlicher Selbstverständnisse im wissenschaftlich-technischen Fortschritt. Freiburg im Breisgau, 216–239.

Grunwald, A. (2019): Technology Assessment in Practice and Theory. Abingdon.

Grunwald, A. (2014): Modes of orientation provided by futures studies: making sense of diversity and divergence. In: European Journal of Futures Research, 2, Art.-Nr. 30 (DOI 10.1007/s40309-013-0030-5).

Grunwald, A. (2007): Technikdeterminismus oder Sozialdeterminismus: Zeitbezüge und Kausalverhältnisse aus der Sicht des „Technology Assessment". In: Dolata, U.; Werle, R. (Hg.): Gesellschaft und die Macht der Technik. Sozioökonomischer und institutioneller Wandel durch Technisierung. Frankfurt am Main, 63–82.

Grunwald, K.; Thiersch, H. (Hg.) (2016): Praxishandbuch Lebensweltorientierte Soziale Arbeit. Handlungszusammenhänge und Methoden in unterschiedlichen Arbeitsfeldern (3. Auflage). Weinheim.

Guckelberger, A. (2019): Öffentliche Verwaltung im Zeitalter der Digitalisierung. Analysen und Strategien zur Verbesserung des E-Governments aus rechtlicher Sicht. Baden-Baden.

Guilford, J. P. (1950): Creativity. In: American Psychologist, 5 (9), 444–454 (DOI: 10.1037/h0063487).

Gutwald, R. et al. (2021): Soziale Konflikte und Digitalisierung. Chancen und Risiken digitaler Technologien bei der Einschätzung von Kindeswohlgefährdungen. In: Ethik Journal, 7 (2). https://www.ethikjournal.de/fileadmin/user_upload/ethikjournal/Texte_Ausgabe_2021_2/Gutwald_u.a._Ethikjournal_2.2021.pdf [16.01.2023].

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

Hacking, I. (1995): The looping effect of human kinds. In: Sperber, D.; Premack, D.; Premack, A. J. (Hg.): Causal Cognition. A Multidisciplinary Debate. Oxford, 351–383.

Hagendorff, T. (2020): The ethics of AI ethics: an evaluation of guidelines. In: Minds and Machines, 30 (1), 99–120 (DOI: 10.1007/s11023-020-09517-8).

Halbig, C. (2007): Praktische Gründe und die Realität der Moral. Frankfurt am Main.

Hansen, J. et al. (2020): Verhaltenskodex für Trusted Learning Analytics. Entwurf für die hessischen Hochschulen. Frankfurt am Main.

Härtel, I. (2019): Digitalisierung im Lichte des Verfassungsrechts – Algorithmen, Predictive Policing, autonomes Fahren. In: Landes- und Kommunalverwaltung, 29 (2), 49–60.

Hartong, S. (2019): Learning Analytics und Big Data in der Bildung. Zur notwendigen Entwicklung eines datenpolitischen Alternativprogramms. Frankfurt am Main.

He, Q.; Turel, O.; Bechara, A. (2017): Brain anatomy alterations associated with Social Networking Site (SNS) addiction. In: Scientific Reports, 7, Art.-Nr. 45064 (DOI: 10.1038/srep45064).

Health Education England (2019): The Topol Review. Preparing the healthcare workforce to deliver the digital future. https://topol.hee.nhs.uk/wp-content/uploads/HEE-Topol-Review-2019.pdf [22.02.2023].

Heckmann, H.-D.; Walter, S. (Hg.) (2006): Qualia. Ausgewählte Beiträge. Paderborn.

Heesen, J. et al. (2020): Zertifizierung von KI-Systemen. Kompass für die Entwicklung und Anwendung vertrauenswürdiger KI-Systeme. https://www.plattform-lernende-systeme.de/files/Downloads/Publikationen/AG1_3_Whitepaper_Zertifizierung_KI_Systemen.pdf [07.02.2023].

Hesse, J. J.; Ellwein, T. (1992): Das Regierungssystem der Bundesrepublik Deutschland (7. Auflage). Opladen.

Hessischer Landtag (2019): Zwischenbericht des Untersuchungsausschusses 19/3. Drucksache 19/6864. https://starweb.hessen.de/cache/DRS/19/4/06864.pdf [01.03.2023].

Heßler, M.; Liggieri, K. (Hg.) (2020): Technikanthropologie. Handbuch für Wissenschaft und Studium. Baden-Baden.

Hevelke, A.; Nida-Rümelin, J. (2015): Responsibility for crashes of autonomous vehicles: an ethical analysis. In: Science and Engineering Ethics, 21 (3), 619–630.

Higgins, M. K. (2021): Can we AlphaFold our way out of the next pandemic? In: Journal of Molecular Biology, 433 (20), Art.-Nr. 167093 (DOI: 10.1016/j.jmb.2021.167093).

Hildebrandt, M. (2022): The issue of proxies and choice architectures. Why EU law matters for recommender systems. In: Frontiers in Artificial Intelligence, 5, Art.-Nr. 789076 (DOI: 10.3389/frai.2022.789076).

Hilgendorf, E. (2020): Digitalisierung, Virtualisierung und das Recht. In: Kasprowicz, D.; Rieger, S. (Hg.): Handbuch Virtualität. Wiesbaden, 405–424.

Hilgendorf, E. (Hg.) (2014): Robotik im Kontext von Recht und Moral. Baden-Baden.

Hochrangige Expertengruppe für künstliche Intelligenz (2019): Ethik-Leitlinien für eine vertrauenswürdige KI. https://data.europa.eu/doi/10.2759/22710 [18.01.2023].

Hoffmann-Riem, W. (2022): Recht im Sog der digitalen Transformation. Tübingen.

Holohan, M.; Fiske, A. (2021): "Like I'm talking to a real person": exploring the meaning of transference for the use and design of AI-based applications in psychotherapy. In: Frontiers in Psychology, 12, Art.-Nr. 720476 (DOI: 10.3389/fpsyg.2021.720476).

Holthausen, J. (2021): Big Data, People Analytics, KI und Gestaltung von Betriebsvereinbarungen – Grund-, arbeits- und datenschutzrechtliche An- und Herausforderungen. In: Recht der Arbeit, 74 (1), 19–32.

Horn, J. L.; Cattell, R. B. (1966): Refinement and test of the theory of fluid and crystallized general intelligences. In: Journal of Educational Psychology, 57 (5), 253–270 (DOI: 10.1037/h0023816).

Horn, C.; Löhrer, G. (Hg.) (2010): Gründe und Zwecke. Texte zur aktuellen Handlungstheorie. Berlin.

Huber, L. (2021): Das rationale Tier. Eine kognitionsbiologische Spurensuche. Berlin.

Hubig, C. (2006): Die Kunst des Möglichen. Band 1: Technikphilosophie als Reflexion der Medialität. Bielefeld.

Hunt, E.; Lunneborg, C.; Lewis, J. (1975): What does it mean to be high verbal? In: Cognitive Psychology, 7 (2), 194–227 (DOI: 10.1016/0010-0285(75)90010-9).

IEEE (2019): Ethically Aligned Design. A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems. https://ieeexplore.ieee.org/document/9398613 [18.01.2023].

Ihde, D. (1990): Technology and the Lifeworld. From Garden to Earth. Bloomington (IN).

Jobin, A.; Ienca, M.; Vayena, E. (2019): The global landscape of AI ethics guidelines. In: Nature Machine Intelligence, 1 (9), 389–399 (DOI: 10.1038/s42256-019-0088-2).

Johnson, D. G. (1985): Computer Ethics. Englewood Cliffs (NJ).

Kaiser, B. (2020): Die Datendiktatur. Wie Wahlen manipuliert werden. Hamburg.

Kaissis, G. et al. (2021): End-to-end privacy preserving deep learning on multi-institutional medical imaging. In: Nature Machine Intelligence, 3 (6), 473–484 (DOI: 10.1038/s42256-021-00337-8).

Kane, R. (Hg.) (2011): The Oxford Handbook of Free Will. New York (NY).

Kant, I. (1968a): Kants Werke. Band 3: Kritik der reinen Vernunft. Berlin.

Kant, I. (1968b): Kants Werke. Band 5: Kritik der praktischen Vernunft / Kritik der Urtheilskraft. Berlin.

Kaspar, J.; Höffler, K.; Harrendorf, S. (2020): Datenbanken, Online-Votings und künstliche Intelligenz – Perspektiven evidenzbasierter Strafzumessung im Zeitalter von „Legal Tech". In: Neue Kriminalpolitik, 32 (1), 35–56 (DOI: 10.5771/0934-9200-2020-1-35).

Kemp, S. (2023): Digital 2023: Global Overview Report. https://datareportal.com/reports/digital-2023-global-overview-report [10.03.2023].

Kettemann, M. C. (2022): UNESCO-Empfehlung zur Ethik Künstlicher Intelligenz. Bedingungen zur Implementierung in Deutschland. https://www.unesco.de/sites/default/files/2022-03/DUK_Broschuere_KI-Empfehlung_DS_web_final.pdf [19.01.2023].

Kiefer, J.; May, M. S. (2022): Diagnostic accuracy and analysis of an artificial intelligence algorithm for the detection of intracranial haemorrhage. In: Insights into Imaging, 13 (Suppl. 4: ECR 2022 Book of Abstracts), 182 (DOI: 10.1186/s13244-022-01337-x).

Knobloch, T. (2018): Vor die Lage kommen: Predictive Policing in Deutschland. Chancen und Gefahren datenanalytischer Prognosetechnik und Empfehlungen für den Einsatz in der Polizeiarbeit. https://www.stiftung-nv.de/sites/default/files/predictive.policing.pdf [10.02.2023].

Kollmer, N.; Klindt, T.; Schucht, C. (Hg.) (2021): Arbeitsschutzgesetz. Kommentar (4. Auflage). München.

Korsgaard, C. M. (1996): The Sources of Normativity. Cambridge.

Kosinski, M.; Stillwell, D.; Graepel, T. (2013): Private traits and attributes are predictable from digital records of human behavior. In: PNAS, 110 (15), 5802–5805 (DOI: 10.1073/pnas.1218772110).

Krafft, T. D.; Gamer, M; Zweig, K. A. (2018): Wer sieht was? Personalisierung, Regionalisierung und die Frage nach der Filterblase in Googles Suchmaschine. Kaiserslautern. https://algorithmwatch.org/de/filterblase-geplatzt-kaum-raum-fuer-personalisierung-bei-google-suchen-zur-bundestagswahl-2017 [12.01.2023].

Kruse, A.; Schmitt, E. (2011): Die Ausbildung und Verwirklichung kreativer Potenziale im Alter im Kontext individueller und gesellschaftlicher Entwicklung. In: Kruse, A. (Hg.): Kreativität im Alter. Heidelberg, 15–46.

Kuhlmann, S.; Trute, H.-H. (2021): Predictive Policing als Formen polizeilicher Wissensgenerierung. In: Zeitschrift für das Gesamte Sicherheitsrecht, 4 (3), 103–110.

Kumkar, L. K. (2022): Plattform-Recht revisited: Umgang mit den Marktordnungen digitaler Plattformen de lege lata et ferenda. In: Zeitschrift für Europäisches Privatrecht, 30 (3), 530–564.

Kumkar, L. K.; Roth-Isigkeit, D. (2020): Erklärungspflichten bei automatisierten Datenverarbeitungen nach der DSGVO. In: Juristenzeitung, 75 (6), 277–286 (DOI: 10.1628/jz-2020-0090).

Kutscher, N. et al. (Hg.) (2020): Handbuch Soziale Arbeit und Digitalisierung. Weinheim.

Landesanstalt für Medien NRW (2022): Hate Speech forsa-Studie 2022. Zentrale Untersuchungsergebnisse. https://www.medienanstalt-nrw.de/fileadmin/user_upload/NeueWebsite_0120/Themen/Hass/LFM_Hatespeech_forsa_2022.pdf [12.01.2023].

Landesanstalt für Medien NRW (2020): forsa-Befragung zu: Hate Speech 2020. Ergebnisbericht. https://www.medienanstalt-nrw.de/fileadmin/user_upload/NeueWebsite_0120/Themen/Hass/forsa_LFMNRW_Hassrede2020_Ergebnisbericht.pdf [12.01.2023].

Landeskriminalamt NRW (2018): Abschlussbericht Projekt SKALA. https://polizei.nrw/sites/default/files/2018-07/180628_Abschlussbericht_SKALA.PDF [17.01.2023].

Laney, D. (2001): 3D Data Management: Controlling Data Volume, Velocity, and Variety. https://studylib.net/doc/8647594 [20.12.2022].

Latour, B. (2007): Eine neue Soziologie für eine neue Gesellschaft. Einführung in die Akteur-Netzwerk-Theorie. Frankfurt am Main.

Laue, P. (2016): Öffnungsklauseln in der DS-GVO – Öffnung wohin? In: Zeitschrift für Datenschutz, 6 (10), 463–467.

Law, J.; Hassard, J. (Hg.) (1999): Actor Network Theory and After. Oxford.

Ley, T.; Reichmann, U. (2020): Digitale Dokumentation in Organisationen Sozialer Arbeit. In: Kutscher, N. et al. (Hg.): Handbuch Soziale Arbeit und Digitalisierung. Weinheim, 241–254.

Libet, B. (2004): Mind Time. The Temporal Factor in Consciousness. Cambridge (MA).

Liesching, M. et al. (2021): Das NetzDG in der praktischen Anwendung. Eine Teilevaluation des Netzwerkdurchsetzungsgesetzes. Berlin.

Lighthill, J. (1973): Artificial intelligence: a general survey. In: Artificial Intelligence. A Paper Symposium. London, 1–21.

van der Linden, S. (2015): The conspiracy-effect: exposure to conspiracy theories (about global warming) decreases pro-social behavior and science acceptance. In: Personality and Individual Differences, 87, 171–173 (DOI: 10.1016/j.paid.2015.07.045).

Lob-Hüdepohl, A. (2021): Messen welcher Wirkung? Normativ-handlungstheoretische Vorbemerkungen zur Wirkungsmessung sozialprofessioneller Interventionen. In: Eurich, J.; Lob-Hüdepohl, A. (Hg.): Gute Assistenz für Menschen in Behinderungen. Wirkungskontrolle und die Frage nach dem gelingenden Leben. Stuttgart, 70–86.

Lob-Hüdepohl, A. (2002): Verantwortung im Verwaltungshandeln. In: Deutsche Verwaltungspraxis, 53 (2), 45–52.

Loh, J. (2017): Strukturen und Relata der Verantwortung. In: Heidbrink, L.; Langbehn, C.; Loh, J. (Hg.): Handbuch Verantwortung. Wiesbaden, 35–56.

Lubart, T. (Hg.) (2018): The Creative Process. Perspectives from Multiple Domains. London.

Lüdemann, C.; Ohlemacher, T. (2002): Soziologie der Kriminalität. Theoretische und empirische Perspektiven. München.

Mainzer, K. (1995): Computer – Neue Flügel des Geistes? Die Evolution computergestützter Technik, Wissenschaft, Kultur und Philosophie. Berlin.

Mann, M. (2020): Technological politics of automated welfare surveillance: social (and data) justice through critical qualitative inquiry. In: Global Perspectives, 1 (1), Art.-Nr. 12991 (DOI: 10.1525/gp.2020.12991).

Martinez-Maldonado, R. et al. (2022): Moodoo the tracker: spatial classroom analytics for characterising teachers' pedagogical approaches. In: International Journal of Artificial Intelligence in Education, 32 (4), 1025–1051 (DOI: 10.1007/s40593-021-00276-w).

Martini, M. et al. (2020): Automatisch erlaubt? Fünf Anwendungsfälle algorithmischer Systeme auf dem juristischen Prüfstand. Gütersloh.

Mast, T. (2023): AGB-Recht als Regulierungsrecht. In: Juristenzeitung, 78 (7), 287–296.

Mattern, F.; Flörkemeier, C. (2010): Vom Internet der Computer zum Internet der Dinge. In: Informatik Spektrum, 33 (2), 107–121 (DOI: 10.1007/s00287-010-0417-7).

Matz, S. C.; Appel, R. E.; Kosinski, M. (2020): Privacy in the age of psychological targeting. In: Current Opinion in Psychology, 31, 116–121 (DOI: 10.1016/j.copsyc.2019.08.010).

Mayor, A. (2018): Gods and Robots. Myths, Machines, and Ancient Dreams of Technology. Princeton (NJ).

McCarthy, J. et al. (2006): A proposal for the Dartmouth Summer Research Project on Artificial Intelligence. In: AI Magazine, 27 (4), 12–14 (DOI: 10.1609/aimag.v27i4.1904).

McDowell, J (2002b): Interne und externe Gründe. In: ders.: Wert und Wirklichkeit. Aufsätze zur Moralphilosophie. Frankfurt am Main, 156–178.

McDowell, J. (2002a): Zwei Arten von Naturalismus. In: ders.: Wert und Wirklichkeit. Aufsätze zur Moralphilosophie. Frankfurt am Main, 30–73.

McDowell, J. (2001): Geist und Welt. Frankfurt am Main.

McKinney, S. M. et al. (2020): International evaluation of an AI system for breast cancer screening. In: Nature, 577 (7788), 89–94 (DOI: 10.1038/s41586-019-1799-6).

Mehta, A. et al. (2021): Acceptability and effectiveness of artificial intelligence therapy for anxiety and depression (Youper): longitudinal observational study. In: Journal of Medical Internet Research, 23 (6), Art.-Nr. e26771 (DOI: 10.2196/26771).

Merleau-Ponty, M. (1966): Phänomenologie der Wahrnehmung. Berlin.

Metzinger, T. (Hg.) (2009): Grundkurs Philosophie des Geistes. Band 1: Phänomenales Bewusstsein. Paderborn.

Meyer-Drawe, K. (2001): Leiblichkeit und Sozialität. Phänomenologische Beiträge zu einer pädagogischen Theorie der Inter-Subjektivität. München.

Mickes, L. et al. (2013): Major memory for microblogs. In: Memory & Cognition, 41 (4), 481–489 (DOI: 10.3758/s13421-012-0281-6).

Miller, T. (2019): Explanation in artificial intelligence: insights from the social sciences. In: Artificial Intelligence, 267, 1–38 (DOI: 10.1016/j.artint.2018.07.007).

Moch, M. (2018): Hilfen zur Erziehung. In: Otto, H.-U. et al. (Hg.): Handbuch Soziale Arbeit (6. Auflage). München, 632–645.

Mohammad, S. M. (2022): Ethics sheet for automatic emotion recognition and sentiment analysis. In: Computational Linguistics, 48 (2), 239–278 (DOI: 10.1162/coli_a_00433).

Molenaar, I. (2021): Personalisation of learning: towards hybrid human-AI learning technologies. In: OECD Digital Education Outlook 2021. Pushing the Frontiers with Artificial Intelligence, Blockchain and Robots. Paris, 57–77.

Moor, J. H. (1985): What is computer ethics? In: Metaphilosophy, 16 (4), 266–275 (DOI: 10.1111/j.1467-9973.1985.tb00173.x).

Muirhead, R.; Rosenblum, N. L. (2019): A Lot of People Are Saying. The New Conspiracism and the Assault on Democracy. Princeton (NJ).

Murphy, G. et al. (2019): False memories for fake news during Ireland's abortion referendum. In: Psychological Science, 30 (10), 1449–1459 (DOI: 10.1177/0956797619864887).

Nagel, T. (1981): Wie ist es, eine Fledermaus zu sein? In: Bieri, P. (Hg.): Analytische Philosophie des Geistes. Königstein im Taunus, 261–275.

Nationale Akademie der Wissenschaften Leopoldina (2021): Digitalisierung und Demokratie. Halle (Saale).

Neftci, E. O.; Averbeck, B. B. (2019): Reinforcement learning in artificial and biological systems. In: Nature Machine Intelligence, 1 (3), 133–143 (DOI: 10.1038/s42256-019-0025-4).

Nguyen, C. T. (2021): How Twitter gamifies communication. In: Lackey, J. (Hg.): Applied Epistemology. Oxford, 410–436.

Nida-Rümelin, J. (2022a): Überblick über die Verwendung der Begriffe starke & schwache Künstliche Intelligenz. In: Chibanguza, K.; Kuß, C.; Steege, H. (Hg.): Künstliche Intelligenz. Recht und Praxis automatisierter und autonomer Systeme. Baden-Baden, 75–90.

Nida-Rümelin, J. (2022b): Digitaler Humanismus – philosophische Aspekte Künstlicher Intelligenz. In: Chibanguza, K.; Kuß, C.; Steege, H. (Hg.): Künstliche Intelligenz. Recht und Praxis automatisierter und autonomer Systeme. Baden-Baden, 29–40.

Nida-Rümelin, J. (2020a): Eine Theorie praktischer Vernunft. Berlin.

Nida-Rümelin, J. (2020b): Die gefährdete Rationalität der Demokratie. Ein politischer Traktat. Hamburg.

Nida-Rümelin, J. (2011): Verantwortung. Stuttgart.

Nida-Rümelin, J. (2005): Über menschliche Freiheit. Stuttgart.

Nida-Rümelin, J.; Weidenfeld, N. (2018): Digitaler Humanismus. Eine Ethik für das Zeitalter der Künstlichen Intelligenz. München.

Nida-Rümelin, J.; Schulenburg, J.; Rath, B. (2012): Risikoethik. Berlin.

Nilsson, N. J. (2010): The Quest for Artificial Intelligence. A History of Ideas and Achievements. Cambridge.

Nkambou, R.; Bourdeau, J.; Mizoguchi, R. (2010): Introduction: what are intelligent tutoring systems, and why this book? In: dies.: Advances in Intelligent Tutoring Systems. Berlin, 1–12.

Obermeyer, Z. et al. (2019): Dissecting racial bias in an algorithm used to manage the health of populations. In: Science, 366 (6464), 447–453 (DOI: 10.1126/science.aax2342).

OECD (2021): OECD Digital Education Outlook 2021. Pushing the Frontiers with Artificial Intelligence, Blockchain and Robots. Paris.

OECD (2019): Recommendation of the Council on Artificial Intelligence. OECD/LEGAL/0449. https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449 [19.01.2023].

Ortiz, J. et al. (2019): Giving voice to the voiceless: the use of digital technologies by marginalized groups. In: Communications of the Association for Information Systems, 45, 20–38 (DOI: 10.17705/1CAIS.04502).

Orwat, C. (2019): Diskriminierungsrisiken durch Verwendung von Algorithmen. Herausgegeben von der Antidiskriminierungsstelle des Bundes. Berlin.

Orwat, C. et al. (2010): Software als Institution und ihre Gestaltbarkeit. In: Informatik Spektrum, 33 (6), 626–633.

Oswald, M. et al. (2018): Algorithmic risk assessment policing models: lessons from the Durham HART model and 'experimental' proportionality. In: Information & Communications Technology Law, 27 (2), 223–250 (DOI:10.1080/13600834.2018.1458455).

Panoptykon Foundation (2015): Profiling the Unemployed in Poland. Social and Political Implications of Algorithmic Decision Making. https://panoptykon.org/sites/default/files/leadimage-biblioteka/panoptykon_profiling_report_final.pdf [30.01.2023].

Pariser, E. (2011): The Filter Bubble. What the Internet Is Hiding from You. New York (NY).

Parviainen, J.; Rantala, J. (2022): Chatbot breakthrough in the 2020s? An ethical reflection on the trend of automated consultations in health care. In: Medicine, Health Care and Philosophy, 25 (1), 61–71 (DOI: 10.1007/s11019-021-10049-w).

Penney, J. W. (2017): Internet surveillance, regulation, and chilling effects online: a comparative case study. In: Internet Policy Review, 6 (2) (DOI: 10.14763/2017.2.692).

Pennycook., G.; Cannon, T. D.; Rand, D. G (2018): Prior exposure increases perceived accuracy of fake news. In: Journal of Experimental Psychology: General, 147 (12), 1865–1880 (DOI: 10.1037/xge0000465).

Petermann, F. (Hg.) (2012): WAIS-IV. Wechsler Adult Intelligence Scale – Fourth Edition. Manual 1: Grundlagen, Testauswertung und Interpretation. Frankfurt am Main.

Polaschek, D. L. L. (2012): An appraisal of the risk-need-responsivity (RNR) model of offender rehabilitation and its application in correctional treatment. In: Legal and Criminological Psychology, 17 (1), 1–17 (DOI: 10.1111/j.2044-8333.2011.02038.x).

Polger, T. W.; Shapiro, L. A. (2016): The Multiple Realization Book. Oxford.

Povalej, R.; Volkmann, D. (2021): Predictive Policing. In: Informatik Spektrum, 44 (1), 57–61 (DOI: 10.1007/s00287-021-01332-4).

Pretschner, A. et al. (2023): Die mächtigen neuen Assistenzsysteme. Was aus der Künstlichen Intelligenz ChatGPT folgt, über die gerade alle sprechen. In: Frankfurter Allgemeine Zeitung, 09.01.2023, S. 18.

Pugh, A. J. (2018): Automated health care offers freedom from shame, but is it what patients need? https://www.newyorker.com/tech/annals-of-technology/automated-health-care-offers-freedom-from-shame-but-is-it-what-patients-need [28.02.2023].

van der Put, C. E. et al. (2016): Detection of unsafety in families with parental and/or child developmental problems at the start of family support. In: BMC Psychiatry, 16, Art.-Nr. 15 (DOI: 10.1186/s12888-016-0715-y).

Putnam, H. (1992): Renewing Philosophy. Cambridge (MA).

Putnam, H. (1975): Philosophy and our mental life. In: ders.: Philosophical Papers. Volume 2: Mind, Language and Reality. Cambridge, 291–303.

Putnam, H. (1967): Psychological predicates. In: Capitan, W. H.; Merrill, D. D. (Hg.): Art, Mind, and Religion. Pittsburgh (PA), 37–48.

Putnam, H. (1965): Brains and behaviour. In: Butler, R. J. (Hg.): Analytical Philosophy. Second Series. Oxford, 1–19.

Quante, M. (2007): Person. Berlin.

Rachels, J. R.; Rockinson-Szapkiw, A. J. (2018): The effects of a mobile gamification app on elementary students' Spanish achievement and self-efficacy. In: Computer Assisted Language Learning, 31 (1–2), 72–89 (DOI: 10.1080/09588221.2017.1382536).

Rademacher, T. (2017): Predictive Policing im deutschen Polizeirecht. In: Archiv des öffentlichen Rechts, 142 (3), 366–416 (DOI: 10.1628/000389117X15054009148798).

Rammert, W.; Schulz-Schaeffer, I. (2002): Technik und Handeln. Wenn soziales Handeln sich auf menschliches Verhalten und technische Artefakte verteilt. In: dies. (Hg.): Können Maschinen handeln? Soziologische Beiträge zum Verhältnis von Mensch und Technik. Frankfurt am Main, 11–64.

Rapp, F. (1978): Analytische Technikphilosophie. Freiburg im Breisgau.

Rau, J. P.; Stier, S. (2019): Die Echokammer-Hypothese: Fragmentierung der Öffentlichkeit und politische Polarisierung durch digitale Medien? In: Zeitschrift für vergleichende Politikwissenschaft, 13 (3), 399–417 (DOI: 10.1007/s12286-019-00429-1).

Raue, B. (2018): Meinungsfreiheit in sozialen Netzwerken. Ansprüche von Nutzern sozialer Netzwerke gegen die Löschung ihrer Beiträge. In: Juristenzeitung, 73 (20), 961–970 (DOI: 10.1628/jz-2018-0261).

Reinholz, D. L.; Stone-Johnstone, A.; Shah, N. (2020): Walking the walk: using classroom analytics to support instructors to address implicit bias in teaching. In: International Journal for Academic Development, 25 (3), 259–272 (DOI: 10.1080/1360144X.2019.1692211).

Rettenberger, M. (2018): Intuitive, klinisch-idiographische und statistische Kriminalprognosen im Vergleich – die Überlegenheit wissenschaftlich strukturierten Vorgehens. In: Forensische Psychiatrie, Psychologie, Kriminologie, 12 (1), 28–36 (DOI: 10.1007/s11757-017-0463-y).

Ribeiro, M. H. et al. (2020): Auditing radicalization pathways on YouTube. In: FAT* '20 Proceedings, 131–141 (DOI: 10.1145/3351095.3372879).

Riehm, T. (2020): Nein zur ePerson! Gegen die Anerkennung einer digitalen Rechtspersönlichkeit. In: Recht Digital, 1 (1), 42–48.

Rip, A. (2007): Die Verzahnung von technologischen und sozialen Determinismen und die Ambivalenzen von Handlungsträgerschaft im „Constructive Technology Assessment". In: Dolata, U.; Werle, R. (Hg.): Gesellschaft und die Macht der Technik. Sozioökonomischer und institutioneller Wandel durch Technisierung. Frankfurt am Main, 83–106.

Rip, A.; Misa, T. J.; Schot, J. (Hg.) (1995): Managing Technology in Society: The Approach of Constructive Technology Assessment. London.

Robb, M. B. (2020): Teens and the News. The Influencers, Celebrities, and Platforms They Say Matter Most. https://www.commonsensemedia.org/sites/default/files/research/report/2020_teensandnews-fullreport_final-release-web.pdf [12.01.2023].

Ropohl, G. (1982): Zur Kritik des technologischen Determinismus. In: Rapp, F.; Durbin, P. T. (Hg.): Technikphilosophie in der Diskussion. Braunschweig, 3–17.

Ross, W. D. (1930): The Right and the Good. Oxford.

Rostalski, F. (2019): Brave New World. Vom (Alp-)Traum lückenloser Straftatenahndung in Zeiten der Digitalisierung. In: Goltdammer's Archiv für Strafrecht, 166 (8), 481–488.

Roxin, C.; Greco, L. (2020): Strafrecht – Allgemeiner Teil. Band I: Grundlagen / Der Aufbau der Verbrechenslehre (5. Auflage). München.

Rudschies, C.; Schneider, I.; Simon, J. (2021): Value pluralism in the AI ethics debate – different actors, different priorities. In: The International Review of Information Ethics, 29, 1–15 (DOI: 10.29173/irie419).

Ryle, G. (1949): The Concept of Mind. London.

Sachverständigenrat zur Begutachtung der Entwicklung im Gesundheitswesen (2021): Digitalisierung für Gesundheit. Ziele und Rahmenbedingungen eines dynamisch lernenden Gesundheitssystems. Bern.

Safdar, N. M.; Banja, J. D.; Meltzer, C. C. (2020): Ethical considerations in artificial intelligence. In: European Journal of Radiology, 122, Art.-Nr. 108768 (DOI: 10.1016/j.ejrad.2019.108768).

Samek, W. et al. (2021): Explaining deep neural networks and beyond: a review of methods and applications. In: Proceedings of the IEEE, 109 (3), 247–278 (DOI: 10.1109/JPROC.2021.3060483).

Samuel, A. L. (1959): Some studies in machine learning using the game of checkers. In: IBM Journal of Research and Development, 3 (3), 210–229 (DOI: 10.1147/rd.33.0210).

Sarsam, S. M. et al. (2020): Sarcasm detection using machine learning algorithms in Twitter: a systematic review. In: International Journal of Market Research, 62 (5), 578–598 (DOI: 10.1177/1470785320921779).

Satterthwaite, M. A. (1975): Strategy-proofness and arrow's conditions: existence and correspondence theorems for voting procedures and social welfare functions. In: Journal of Economic Theory, 10 (2), 187–217 (DOI: 10.1016/0022-0531(75)90050-2).

von Savigny, E. (1974): Die Philosophie der normalen Sprache. Eine kritische Einführung in die „ordinary language philosophy". Frankfurt am Main.

Scanlon, T. M (2014): Being Realistic about Reasons. Oxford.

Scanlon, T. M (1998): What We Owe to Each Other. Cambridge (MA).

Schelb, P. et al. (2019): Classification of cancer at prostate MRI: deep learning versus clinical PI-RADS assessment. In: Radiology, 293 (3), 607–617 (DOI: 10.1148/radiol.2019190938).

Schlemmer, H.-P.; Hohenfellner, M. (2021): Chancen von KI in der Onkologie am Beispiel der individualisierten Diagnostik und Behandlung von Prostatakrebs. In: Zeitschrift für medizinische Ethik, 67 (3), 309–326.

Schlick, M. (1930): Fragen der Ethik. Wien.

Schmid, U. (2022): Vertrauenswürdige Künstliche Intelligenz. Nachvollziehbar, Transparent, Korrigierbar. In: Bundesministerium für Umwelt, Naturschutz, nukleare Sicherheit und Verbraucherschutz; Rostalski, F. (Hg.): Künstliche Intelligenz. Wie gelingt eine vertrauenswürdige Verwendung in Deutschland und Europa? Tübingen, 287–297.

Schmidt, M. G. (2010): Wörterbuch zur Politik (3. Auflage). Stuttgart.

Schmidt-Aßmann, E. (1998): Das allgemeine Verwaltungsrecht als Ordnungsidee. Grundlagen und Aufgaben der verwaltungsrechtlichen Systembildung. Berlin.

Schmitz, H. (1990): Der unerschöpfliche Gegenstand. Grundzüge der Philosophie. Bonn.

Schneider, D. (2020): Decision Support Systeme in der Sozialen Arbeit – Herausforderungen an die Rolle der TA in Innovationsprozessen. In: Nierling, L.; Torgersen, H. (Hg.): Die neutrale Normativität der Technikfolgenabschätzung. Konzeptionelle Auseinandersetzung und praktischer Umgang. Baden-Baden, 117–138.

Schrittwieser, J. et al. (2020): Mastering Atari, Go, chess and shogi by planning with a learned model. In: Nature, 588 (7839), 604–609 (DOI: 10.1038/s41586-020-03051-4).

Schrödter, M.; Bastian, P.; Taylor, B. (2020): Risikodiagnostik und Big Data Analytics in der Sozialen Arbeit. In: Kutscher, N. et al. (Hg.): Handbuch Soziale Arbeit und Digitalisierung. Weinheim, 255–264.

Schulz, W.; Schmees, J. (2022): Möglichkeiten und Grenzen der Künstlichen Intelligenz in der Rechtsanwendung. In: Augsberg, I.; Schuppert, G. F. (Hg.): Wissen und Recht. Baden-Baden, 561–594.

Schumpeter, J. A. (2018): Kapitalismus, Sozialismus und Demokratie. Tübingen.

Schwabe, M. (2022): Die „dunklen Seiten" der Sozialpädagogik. Über den Umgang mit Fehlern, Unvermögen, Ungewissheit, Ambivalenzen, Idealen und Destruktivität. Weinheim.

Schwabe, M. (2008): Methoden der Hilfeplanung. Zielentwicklung, Moderation und Aushandlung. Frankfurt am Main.

Schwartz, J. T. (1986): The Limits of Artificial Intelligence. Article for 'Encyclopedia of Artificial Intelligence'. New York (NY).

Searle, J. R. (1980): Minds, brains, and programs. In: Behavioral and Brain Sciences, 3 (3), 417–424 (DOI: 10.1017/S0140525X00005756).

Sechopoulos, I.; Teuwen, J.; Mann, R. (2021): Artificial intelligence for breast cancer detection in mammography and digital breast tomosynthesis: state of the art. In: Seminars in Cancer Biology, 72, 214–225 (DOI: 10.1016/j.semcancer.2020.06.002).

Shaw, S. W. (1990): Pattern recognition. In: AI Magazine, 11 (2), 80–81 (DOI: 10.1609/aimag.v11i2.838).

Simitis, S.; Hornung, D.; Spiecker gen. Döhmann, I. (Hg.) (2019): Datenschutzrecht. DSGVO mit BDSG (1. Auflage). Baden-Baden.

Singelnstein, T. (2018): Predictive Policing: Algorithmenbasierte Straftatprognosen zur vorausschauenden Kriminalintervention. In: Neue Zeitschrift für Strafrecht, 38 (1), 1–9.

Singer, W. (2004): Verschaltungen legen uns fest: Wir sollten aufhören, von Freiheit zu sprechen. In: Geyer, C. (Hg.): Hirnforschung und Willensfreiheit. Zur Deutung der neuesten Experimente. Frankfurt am Main, 30–65.

Solove, D. J. (2006): A taxonomy of privacy. In: University of Pennsylvania Law Review, 154 (3), 477–564 (DOI: 10.2307/40041279).

Sommerer, L. M. (2020): Personenbezogenes Predictive Policing. Kriminalwissenschaftliche Untersuchung über die Automatisierung der Kriminalprognose. Baden-Baden.

Ständige Wissenschaftliche Kommission der Kultusministerkonferenz (2022): Digitalisierung im Bildungssystem.: Handlungsempfehlungen von der Kita bis zur Hochschule. https://www.kmk.org/fileadmin/Dateien/pdf/KMK/SWK/2022/SWK-2022-Gutachten_Digitalisierung.pdf [10.03.2023].

Stark, B.; Magin, M; Jürgens, P. (2021): Maßlos überschätzt. Ein Überblick über theoretische Annahmen und empirische Befunde zu Filterblasen und Echokammern. In: Eisenegger, M. et al. (Hg.): Digitaler Strukturwandel der Öffentlichkeit. Historische Verortung, Modelle und Konsequenzen. Wiesbaden, 303–321.

Sternberg, R. J. (Hg.) (2020): The Cambridge Handbook of Intelligence. Cambridge.

Stewart, A. J. et al. (2019): Information gerrymandering and undemocratic decisions. In: Nature, 573 (7772), 117–121 (DOI: 10.1038/s41586-019-1507-6).

Strawson, P. F. (1974): Freedom and Resentment and Other Essays. London.

Suchman, L. A. (2007): Human-Machine Reconfigurations. Plans and Situated Actions. Cambridge.

Suchman, L. A. (1987). Plans and Situated Actions. The Problem of Human-Machine Communication. Cambridge.

Tcherkassof, A.; Dupré, D. (2022): The emotion-facial expression link: evidence from human and automatic expression recognition. In: Psychological Research, 85 (8), 2954–2969 (DOI: 10.1007/s00426-020-01448-4).

Thaler, R. H.; Sunstein, C. R. (2011): Nudge. Wie man kluge Entscheidungen anstößt. Berlin.

Thurstone, L. L. (1938): Primary Mental Abilities. Chicago (IL).

Thurstone, L. L.; Thurstone, T. G. (1941): Factorial Studies of Intelligence. Chicago (IL).

Törnberg, P. (2022): How digital media drive affective polarization through partisan sorting. In: PNAS, 119 (42), Art.-Nr. e2207159119 (DOI: 10.1073/pnas.2207159119).

Trappe, T. (2013): Ethik der öffentlichen Verwaltung – eine Skizze. In: Büsch, D.; Kutscha, M. (Hg.): Recht, Lehre und Ethik der öffentlichen Verwaltung. Baden-Baden, 145–162.

Treuthardt, D.; Kröger, M. (2019): Der Risikoorientierte Sanktionenvollzug (ROS) – empirische Überprüfung des Fall-Screening-Tools (FaST). In: Schweizerische Zeitschrift für Kriminologie, 18 (1), 76–85.

Treuthardt, D.; Loewe-Baur, M.; Kröger, M. (2018): Der Risikoorientierte Sanktionenvollzug (ROS) – aktuelle Entwicklungen. In: Schweizerische Zeitschrift für Kriminologie, 17 (2), 24–32.

Truby, J.; Brown, R. (2021): Human digital thought clones: the Holy Grail of artificial intelligence for big data. In: Information & Communications Technology Law, 30 (2), 140–168 (DOI: 10.1080/13600834.2020.1850174).

Turing, A. M. (1950): Computing Machinery and Intelligence. In: Mind, 59 (236), 433–460 (DOI: 10.1093/mind/LIX.236.433).

UNESCO (2021): Recommendation on the Ethics of Artificial Intelligence. SHS/BIO/REC-AIETHICS/2021. https://unesdoc.unesco.org/ark:/48223/pf0000380455 [18.01.2023].

United States Senate (2018): S. Hrg. 115-232 - Open Hearing: Social Media Influence in the 2016 U.S. Election. https://www.govinfo.gov/content/pkg/CHRG-115shrg27398/pdf/CHRG-115shrg27398.pdf [09.03.2023].

Uscinski, J. et al. (2022): Have beliefs in conspiracy theories increased over time? In: PLOS One, 17 (7), Art.-Nr. e0270429 (DOI: 10.1371/journal.pone.0270429).

Van Daele, T. et al. (2021). Dropping the E: the potential for integrating e-mental health in psychotherapy. In: Current Opinion in Psychology, 41, 46–50 (DOI: 10.1016/j.copsyc.2021.02.007).

Vincent-Lancrin, S. (2021): Frontiers of smart education technology: opportunities and challenges. In: OECD Digital Education Outlook 2021. Pushing the Frontiers with Artificial Intelligence, Blockchain and Robots. Paris, 19–42.

Vogt, M.; Leser, U.; Akbik, A. (2021): Early detection of sexual predators in chats. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 4985–4999 (DOI: 10.18653/v1/2021.acl-long.386).

Vosoughi, S.; Roy, D.; Aral, S. (2018): The spread of true and false news online. In: Science, 359 (6380), 1146–1151 (DOI: 10.1126/science.aap9559).

de Vries, H. B.; Lubart, T. I. (2017): Scientific creativity: divergent and convergent thinking and the impact of culture. In: Journal of Creative Behavior, 53 (2), 145–155 (DOI: 10.1002/jocb.184).

Waas, B. (2022): KI und Arbeitsrecht. In: Recht der Arbeit, 75 (3), 125–130.

Wachs, S; Wolf, K. D.; Pan, C.-C. (2012): Cybergrooming: risk factors, coping strategies and associations with cyberbullying. In: Psicothema, 24 (4), 628–633.

Ward, A. F. et al. (2017): Brain drain: the mere presence of one's own smartphone reduces available cognitive capacity. In: Journal of the Association for Consumer Research, 2 (2), 140–154 (DOI: 10.1086/691462).

Warnat-Herresthal, S. et al. (2021): Swarm Learning for decentralized and confidential clinical machine learning. In: Nature, 594 (7862), 265–270 (DOI: 10.1038/s41586-021-03583-3).

Watson, J. B. (1930): Der Behaviorismus. Stuttgart.

Wechsler, D. (1944): The Measurement of Adult Intelligence. Baltimore (MD).

Weiden, H. (2022): Mehr Freiheit und Sicherheit im Netz. Gutachten zum Entwurf des Digital Services Act im Auftrag der Friedrich-Naumann-Stiftung für die Freiheit. https://shop.freiheit.org/#!/Publikation/1201 [08.02.2023].

Weingart, P. (Hg.) (1989): Technik als sozialer Prozess. Frankfurt am Main.

Weizenbaum, J. (1976): Computer Power and Human Reason. From Judgment to Calculation. San Francisco (CA).

Weyer, J. et al. (1997): Technik, die Gesellschaft schafft. Soziale Netzwerke als Ort der Technikgenese. Berlin.

Wielsch, D. (2018): Die Ordnungen der Netzwerke. AGB – Code – Community Standards. In: Eifert, M.; Gostomzyk, T. (Hg.): Netzwerkrecht. Die Zukunft des NetzDG und seine Folgen für die Netzwerkkommunikation. Baden-Baden, 61–94.

Wiener, N. (1950): The Human Use of Human Beings. Cybernetics and Society. Boston (MA).

Wittgenstein, L. (1953): Philosophical Investigations. Oxford.

Wojcik, S. et al. (2018): Bots in the Twittersphere. https://www.pewresearch.org/internet/2018/04/09/bots-in-the-twittersphere [12.01.2023].

Yeung, K. (2017): 'Hypernudge': big data as a mode of regulation by design. In: Information, Communication & Society, 20 (1), 118–136 (DOI: 10.1080/1369118X.2016.1186713).

Yeung, K. (2018): A Study of the Implications of Advanced Digital Technologies (Including AI systems) for the Concept of Responsibility within a Human Rights Framework. https://ssrn.com/abstract=3286027 [10.02.2023].

Zeller, C.; Schmid, U. (2017): Automatic Generation of Analogous Problems to Help Resolving Misconceptions in an Intelligent Tutor System for Written Subtraction. https://fis.uni-bamberg.de/handle/uniba/41882 [11.01.2023].

Zentrale Ethikkommission bei der Bundesärztekammer (2021): Entscheidungsunterstützung ärztlicher Tätigkeit durch Künstliche Intelligenz. https://www.aerzteblatt.de/down.asp?id=28458 [10.02.2023].

ABKÜRZUNGSVERZEICHNIS

Abs.	Absatz
ADM	automated/algorithmic decision making
AI	artificial intelligence
ALLEGRO	Arbeitslosengeld II Leistungsverfahren Grundsicherung Online
AMAS	Arbeitsmarktchancen-Assistenzsystem
AMS	Arbeitsmarktservice
Art.	Artikel
CADe	computer-aided detection
CADx	computer-aided diagnose
COMPAS	Correctional Offender Management Profiling for Alternative Sanctions
COVID-19	coronavirus disease 2019
DSGVO	Datenschutz-Grundverordnung
EEG	Elektroenzephalogramm
EU	Europäische Union
f.	folgende [Seite]
ff.	folgende [Seiten]
FlugDaG	Fluggastdatengesetz
GPT	generative pre-trained transformer
HSOG	Hessischen Gesetzes über die öffentliche Sicherheit und Ordnung
IT	Informationstechnologie
ITS	intelligentes Tutorsystem
KI	Künstliche Intelligenz
mpMRT	multiparametrische Magnetresonanztomografie
MStV	Medienstaatsvertrag
Nr.	Nummer
PSA	prostataspezifisches Antigen
Rn.	Randnummer
ROS	Risikoorientierter Sanktionenvollzug
SGB	Sozialgesetzbuch
StGB	Strafgesetzbuch
USA	United States of America
vgl.	vergleiche

Mitglieder des Deutschen Ethikrates

Prof. Dr. med. Alena Buyx (Vorsitzende)
Prof. Dr. iur. Dr. h. c. Volker Lipp (Stellvertretender Vorsitzender)
Prof. Dr. phil. Dr. h. c. Julian Nida-Rümelin (Stellvertretender Vorsitzender)
Prof. Dr. rer. nat. Susanne Schreiber (Stellvertretende Vorsitzende)

Prof. Dr. iur. Steffen Augsberg
Regionalbischöfin Dr. phil. Petra Bahr
Prof. Dr. theol. Franz-Josef Bormann
Prof. Dr. rer. nat. Hans-Ulrich Demuth
Prof. Dr. iur. Helmut Frister
Prof. Dr. theol. Elisabeth Gräb-Schmidt
Prof. Dr. rer. nat. Dr. phil. Sigrid Graumann
Prof. Dr. rer. nat. Armin Grunwald
Prof. Dr. med. Wolfram Henn
Prof. Dr. rer. nat. Ursula Klingmüller
Stephan Kruip
Prof. Dr. theol. Andreas Lob-Hüdepohl
Prof. Dr. phil. habil. Annette Riedel
Prof. Dr. iur. Stephan Rixen
Prof. Dr. iur. Dr. phil. Frauke Rostalski
Prof. Dr. theol. Kerstin Schlögl-Flierl
Dr. med. Josef Schuster
Prof. Dr. phil. Mark Schweda
Prof. Dr. phil. Judith Simon
Prof. Dr. phil. Muna Tatari

Externer Experte
Prof. Dr. phil. habil. Dr. phil. h. c. lic. phil. Carl Friedrich Gethmann
(Ratsmitglied bis 13. Februar 2021, danach Mitarbeit als externer Experte)