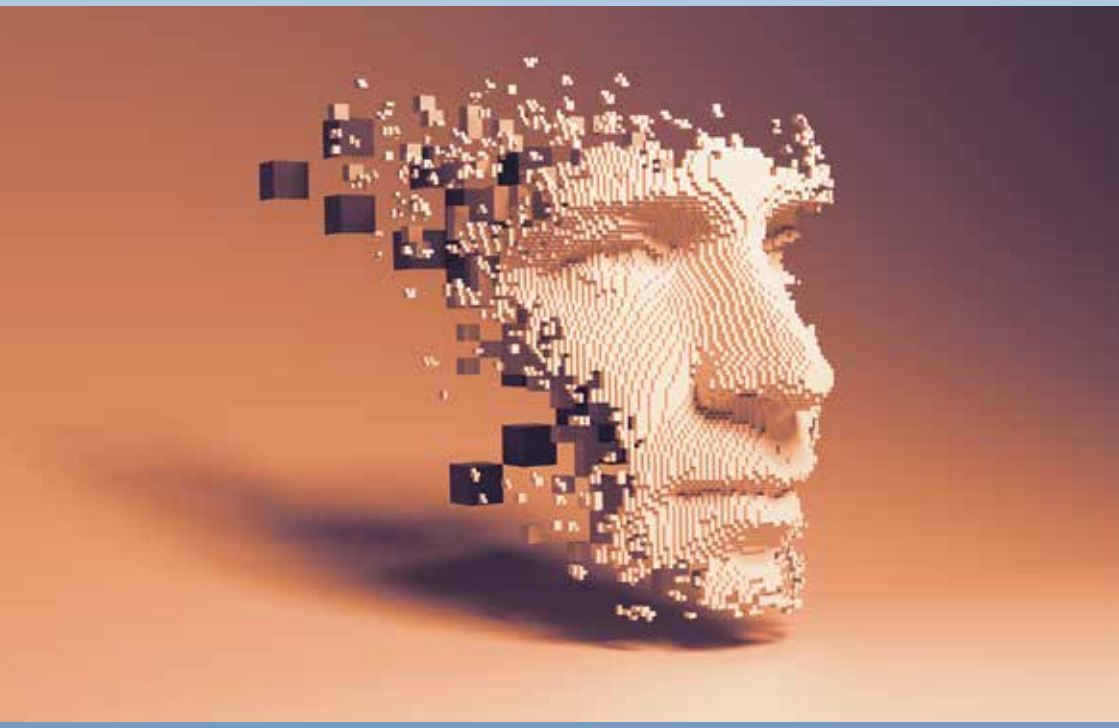
Deutscher Ethikrat



Mensch und Maschine – Herausforderungen durch Künstliche Intelligenz

STELLUNGNAHME · KURZFASSUNG

20. März 2023

Der vollständige Text der Stellungnahme „Mensch und Maschine – Herausforderungen durch Künstliche Intelligenz“ sowie alle öffentlich verfügbaren begleitenden Informationen und Dokumentationen des Deutschen Ethikrates zum Thema sind unter https://www.ethikrat.org/themen/forschung-und-technik/mensch-und-maschine abrufbar.

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

Layout: Torsten Kulick
Titelillustration: pinkeyes/Shutterstock.com

>> INHALT

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“ (2019) angeschnitten wurden. Der Deutsche Ethikrat reagiert mit der Stellungnahme 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, philosophischen und methodischen 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 englische Mathematiker Alan Turing formulierte 1950 ein später als Turing-Test bezeichnetes Kriterium für KI, nachdem maschinelle Intelligenz dann vorläge, wenn das Verhalten einer Maschine für menschliche Beobachter nicht von dem eines Menschen unterscheidbar erscheint.

5) Die frühe Forschung zur 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 vollständig einsehen kann, nicht auf Anhieb 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 algorithmisch gestützte 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 künstlicher 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 und künstlicher Intelligenz 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 leib-seelische 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 auswirkt, 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 einer Handlung beispielsweise neben den beabsichtigten Folgen auch nicht beabsichtigte, aber dem 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 (Betroffene), 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 Freiheitsfähigkeit 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 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 sozialem 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 Randbedingungen und Erfolgsbedingungen für weiteres menschliches Leben setzt und auch 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 Maß 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 von jenen, *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 Entscheidern. 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 vorneherein 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 Tumore 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 Protokollen 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 für einzelne 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 Video- oder Audiodaten 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, müssen 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 es 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 Inhalte 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 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 Informationsqualität und 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 und 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: Ursache vieler informationeller und kommunikativer 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 für verschiedene 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 Straftätern 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 AMAS-System 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 Benutzeroberflächen verschwindet.

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 mit teils überzogenen Hoffnungen, aber auch mit teilweise 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 horizontales wie vertikales, gestaltendes 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 sektorenübergreifende 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 der 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, beeinträchtigt nicht nur die Privatsphäre der Personen, von denen diese Daten stammen, sondern macht 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 bereist 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 der Datengeber 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 der kommunalen 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 mit Blick auf die 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 KI-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, muss 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 zu 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. Die Transparenz und Nachvollziehbarkeit algorithmischer Systeme steht zwar in Zusammenhang mit deren Kontrolle und der Verantwortung für ihren Einsatz, ist 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 in 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.

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)

Mitarbeiterinnen und Mitarbeiter der Geschäftsstelle

Dr. rer. nat. Joachim Vetter (Leiter)
Carola Böhm
Luca Böllert
Ulrike Florian
Dr. phil. Thorsten Galert
Steffen Hering
Petra Hohmann
Jonas Huggins
Torsten Kulick
Maximilian Lübben
Dr. rer. nat. Lilian Marx-Stölting
Dr. Nora Schultz
Anneke Viertel