

Christian Albrecht (Hrsg.)

# KÜNSTLICHE INTELLIGENZ IN DIAKONISCHEN UNTERNEHMEN

VON SCIENCE FICTION ZUR ALLTAGSREALITÄT

Diakonie reflektiert – Band 3 (2025)

Eine Reihe der Evangelischen Arbeitsstelle  
für missionarische Kirchenentwicklung  
und diakonische Profilbildung (midi)

**mid**i

## EDITORIAL

„Bußtagstagungen“ ist die interne, inoffizielle Bezeichnung für die alljährlich stattfindenden Tagungen für Führungskräfte der Diakonie, die regelmäßig in der Evangelischen [Akademie Tutzing](#) stattfinden. Seit zwanzig Jahren laden die drei großen diakonischen Unternehmen in Bayern, das [Augustinum](#), [diakoneo](#) und die [Rummelsberger Diakonie](#) ein; die fachliche Leitung liegt seit vielen Jahren bei Christian Albrecht, Praktischer Theologe an der [LMU München](#). Immer geht es auf den Tagungen darum, grundsätzliche Fragestellungen der alltäglichen diakonischen Arbeit, für deren Reflexion im Alltag häufig nicht genug Zeit und Gelegenheit ist, einmal mit dem nötigen Abstand vom Alltag zu bedenken. Während viele Tagungen der vergangenen Jahre in [Buchform](#) veröffentlicht worden sind, werden sie seit 2023 als Netzpublikation in der Reihe „[Diakonie reflektiert](#)“ dokumentiert.

## INHALT

Editorial.....	2
<i>Christian Albrecht</i>	
Vorwort.....	4
<i>Rüdiger Schuch</i>	
Künstliche Intelligenz in Diakonischen Unternehmen – Eine Einführung.....	8
<i>Theresa Züger</i>	
Wie kann KI im Sinne des Gemeinwohls eingesetzt werden?.....	24
<i>Elisabeth André</i>	
Künstliche Intelligenz und Empathie – passt das zusammen und falls ja, wie?.....	44
<i>Sven Nyholm</i>	
Ethik der KI: Eine philosophische Perspektive.....	75
<i>Christian Albrecht</i>	
Natürliche Intelligenz. Theologisch-Ethische Aspekte des Einsatzes von KI im Christlichen Sozial- und Gesundheitswesen.....	89
Impressum.....	105

*Christian Albrecht*

## VORWORT

An KI scheiden sich die Geister. Die einen erwarten von KI Rationalisierungen, Effizienzsteigerungen, Aufwandssenkungen – die anderen befürchten die schleichende Abschaffung menschlicher Erfindungsgabe, Gestaltungskraft und Autonomie. Den einen gilt KI als Heilsverheißung, den anderen als Apokalypse. Übertriebene Erwartungen stehen unvermittelt neben unmäßigen Ängsten, und vielfach fehlt es an nüchterner Sachkunde zur besonnenen Beurteilung von Chancen und Grenzen der KI.

Besonders herausforderungsvoll ist die realistische Einschätzung der Möglichkeiten und Einschränkungen des Einsatzes von KI im Bereich des Sozial- und Gesundheitswesens. Hier zeigt sich die genannte Spannung verschärft. Einerseits gibt es zahlreiche Anwendungsfelder, auf denen KI zeit- und ressourcenraubende Belastungen des Pflegepersonals, aber auch der Klienten und Klientinnen erheblich reduzieren kann. Andererseits sind die Hemmungen zum pragmatischen Einsatz von KI groß, weil sie die Ideale personaler Zuwendung, zwischenmenschlicher Beziehung und ganzheitlicher Authentizität des individuell Persönlichen zu verdrängen drohen. Aber ist dem wirklich so, auf beiden Seiten? Reduziert KI den Aufwand? Verdrängt KI das unverwechselbar Menschliche?

Gerade im Bereich christlicher Sozialfürsorge müssen wir dringend unser Wissen über KI steigern und zu einer umsichtigen Einschätzung des ethisch Vertretbaren kommen. Was wird längst und geräuschlos praktiziert? Was sind realistische Zukunftserwartungen? Und was sind absurde Szenarien?

Die Überlegungen dieser Netzpublikation, die die Beiträge auf der Bußtagstagung des Jahres 2024 dokumentiert, sollen diesen Fragen, die den Alltag der in der Diakonie Tätigen bestimmen, zugleich praxisnah wie aus dem reflexiven Abstand nachgehen.

Eine Einführung ins Thema liefert [Rüdiger Schuch](#), Präsident der [Diakonie Deutschland](#). Er nennt Gründe dafür, dass KI auch die Diakonie grundlegend verändern wird und beschreibt Chancen und Risiken

dieser Veränderung. Dazu wird der Bedarf an KI in den diakonischen Einrichtungen und Unternehmen im Blick auf konkrete Anwendungen konkretisiert. Vor allem aber geht es ihm um Leitlinien der Anwendung von KI-Tools, die den ethischen Herausforderungen von KI in der Diakonie gerecht werden. Leitend ist dabei der Gedanke, dass KI eine unterstützende Funktion hat. KI darf menschliche Begegnungen nicht ersetzen, ihr Einsatz muss transparent gestaltet werden und dem Schutz sensibler Daten verpflichtet sein. Unter diesen Bedingungen plädiert Schuch für strategische Investitionen der diakonischen Einrichtungen und Unternehmen in KI-Technologien.

Schon aus dieser Einführung geht hervor: Eine der zahlreichen Besonderheiten von KI im christlichen Gesundheits- und Sozialwesen besteht darin, dass KI hier nicht zu einseitiger Nutzenoptimierung eingesetzt werden kann, sondern dass bei diesem Einsatz in ausgleichender Weise verschiedene Interessen unterschiedlicher Beteiligter berücksichtigt werden müssen. In einer grundsätzlichen Weise werden diese Herausforderungen bedacht in Überlegungen zu Theorie und Praxis gemeinwohlorientierten Einsatzes von KI. Darüber informiert [Theresa Züger](#), Leiterin des [AI & Society Lab](#) am Alexander von Humboldt Institut für Internet und Gesellschaft in Berlin. Ausgehend von Gemeinwohltheorien und der Untersuchung von gemeinwohlorientierten KI-Projekten entfaltet sie zentrale Prinzipien des gemeinwohlorientierten Einsatzes: Rechtfertigung des KI-Einsatzes, Gleichberechtigung (Equity), Partizipation, technische Sicherheit, Transparenz und Nachhaltigkeit. Auf dem Hintergrund von Informationen über die technischen Grundlagen von KI (u.a. neuronale Netze und Natural Language Processing) zeigt sie Möglichkeiten von deren Umsetzung in sozialen Kontexten.

Ein zentraler Punkt in den Überlegungen zum Einsatz von KI in sozialen Kontexten besteht in der Frage, ob in der Delegation von Aufgaben und Funktionen an KI-gesteuerte Techniken nicht das Menschliche zu kurz komme, die Zuwendung, das Persönliche, das Empathische. Die KI wäre nicht die KI, wenn es nicht aussichtsreiche Versuche gäbe, auch das digital zu simulieren, und es geschieht unter dem Stichwort der Artificial Empathy. [Elisabeth André](#), Expertin für die Interaktion zwischen Mensch und digitaler Maschine, informiert über den Stand der Forschung zu Artificial Empathy. Sie zeigt zunächst, dass Emotionen eine Grundlage rationaler Entscheidungen sind und beschreibt, wie künstliche Systeme emotionale Signale erkennen und simulieren kön-

nen. Dabei unterscheidet sie verschiedene Formen von Empathie (ideomotorisch, affektiv, kognitiv, funktional) und geht auf die technischen Herausforderungen bei der Emotionserkennung und -darstellung durch Maschinen ein. André warnt vor der Überschätzung maschineller Fähigkeiten, insbesondere im Hinblick auf tiefes emotionales Verstehen und Kontextsensitivität, und nennt mögliche Risiken sowie ethische Fragen im Umgang mit empathisch agierenden KI-Systemen. Gleichwohl stellt sie das große Potenzial emotionaler KI etwa in Kontexten von Bildung und Psychotherapie heraus.

Alle vorangehenden Informationen über Möglichkeiten und Grenzen von KI im christlichen Gesundheits- und Sozialwesen zeigen den hohen ethischen Orientierungsbedarf bei der Entscheidung über konkrete Einsätze. Diesem ethischen Orientierungsbedarf sind die beiden abschließenden Beiträge gewidmet.

Den Anfang macht [Sven Nyholm](#), Professor für Ethik der Künstlichen Intelligenz, mit Überlegungen aus der Sicht philosophischer Ethik. Anhand konkreter Beispiele belegt er, dass herkömmliche, personal orientierte Zuschreibungspraktiken und Verantwortungsmodelle vielfach nicht mehr ausreichen. So entsteht die Notwendigkeit, traditionelle ethische Theorien insbesondere im Hinblick auf neue Akteure, die weder eindeutig moralische Subjekte noch Objekte sind, auszuweiten. Dabei plädiert er für eine differenzierte Betrachtung zwischen technischer (enger) und lebensweltlich orientierter (weiter) Ethik, zwischen negativer (vermeidender) und positiver (gestaltender) Perspektive sowie für neue Prinzipien zur Regulierung von Interaktionen zwischen Mensch und Mensch, zwischen Mensch und KI sowie zwischen KI und KI.

[Christian Albrecht](#) schließt mit Überlegungen zum Einsatz von KI in der Diakonie aus der Sicht theologischer Ethik. Er fragt nach theologischen Kriterien zum Einsatz von KI und findet sie in Gesichtspunkten, die sich aus dem christlichen Schöpfungsgedanken ableiten lassen: der Sicherung menschlicher Freiheit und der Ermöglichung individueller Entwicklung. Aus ihnen ergibt sich eine Art Stufenschematik. Als unproblematisch können KI-Anwendungen gelten, die rein administrative Aufgaben unterstützen. Problematisch, aber unter Umständen vertretbar, sind solche, die in Kommunikation eingreifen, sofern ihre künstliche Natur transparent bleibt. Eher abzulehnen ist der KI-Einsatz, wenn er zu einer Abhängigkeit von intransparenten Entscheidungssystemen führt, etwa in der medizinischen Diagnostik, wo dies das Vertrauen des Patien-

ten zum Arzt tangiert. So plädiert der Vortrag für eine bewusst gestaltete, kriteriengeleitete Nutzung von KI unter Wahrung menschlicher Freiheit und Individualität.

Insgesamt bieten die hier dokumentierten Tagungsbeiträge zahlreiche Anregungen und Anhaltspunkte für einen reflektierten Umgang mit der Ambivalenz von KI in diakonischen Einrichtungen und Unternehmen. Einerseits verbietet es sich, das unterstützende Potential von KI-Anwendungen im christlichen Sozialwesen zu ignorieren: das widerspräche allen Grundsätzen der pragmatischen Vernunft und der Wirkungsorientierung in der Diakonie. Andererseits gibt es Grenzen dessen, was man KI-Tools in der Diakonie überlassen möchte. Diese Grenzen werden vor allem durch ethische Aspekte markiert. Diese ethischen Aspekte müssen aber aufs Konkrete bezogen werden, genauer: auf das – sich permanent verändernde – konkret Realistische. Das bringt für die Verantwortlichen in diakonischen Einrichtungen und Unternehmen neue und dauerhafte Reflexionsaufgaben mit sich. Zu deren Bewältigung möchten die nachfolgenden Überlegungen Orientierungshilfen geben.

Mein Dank für ihre Hilfe bei der Redaktion dieses Bandes gilt den Lehrstuhlmitarbeitern Andreas Eder und Micha Kettling.

*Rüdiger Schuch*

## KÜNSTLICHE INTELLIGENZ IN DIAKONISCHEN UNTERNEHMEN – EINE EINFÜHRUNG

### *Hinführung*

Wir stehen an einem entscheidenden Wendepunkt in der Geschichte der Technologie. Die menschliche Zivilisation hat bereits zahlreiche Wellen technologischer Revolutionen erlebt, die die Gesellschaft tiefgreifend verändert haben: Vom Buchdruck über die Massenmedien hin zur Digitalisierung. Nun stehen wir am Beginn der nächsten Welle: die Ära der Künstlichen Intelligenz und synthetischen Biologie, die unsere Welt grundlegend verändern.<sup>1</sup>

KI ist weit mehr als nur eine neue Technologie, die Maschinen befähigt, menschenähnliche Fähigkeiten zu erlernen. Wir erleben derzeit einen epochalen Medienwandel, der nicht nur die Verbreitung, sondern vor allem die Generierung von Wissen revolutioniert. Die Möglichkeit, alle Wissensbestände miteinander zu verknüpfen, birgt ein enormes Potenzial für Neues, das es so noch nie gegeben hat. Zum ersten Mal haben wir eine Technologie entwickelt, die selbst Inhalte generieren kann, was Auswirkungen mit sich bringt, die schwer zu kontrollieren sind.

Mustafa Suleyman, ein renommierter KI-Experte und Mitbegründer des KI-Unternehmens DeepMind äußert sich dazu mit nachdenklicher Besorgnis: „Mit den Fortschritten, die die Technologie im Laufe der Jahre gemacht hat, sind meine Bedenken gewachsen. Was, wenn die Welle in Wirklichkeit ein Tsunami ist?“<sup>2</sup>

In den vergangenen Jahren hat sich der KI-Boom in rasanter Geschwindigkeit entwickelt. KI-Systeme werden günstiger, leichter zugänglich und verbreiten sich dadurch in weiten Teilen der Gesellschaft. So wird KI zunehmend Teil aller Lebensbereiche unserer Lebens- und Arbeitswelt: Von Kommunikation und Bildung über Biotechnologie und

---

1 Vgl. Mustafa Suleyman: *The coming wave. Künstliche Intelligenz, Macht und das größte Dilemma des 21. Jahrhunderts*, München 2024, S. 13.

2 A.a.O., S. 16.

Cyberkriminalität bis hin zu Fragen der Demokratie und Umweltauswirkungen. Mustafa Suleyman spricht davon, dass KI-Systeme „außergewöhnliche neue medizinische Fortschritte und bahnbrechende Entwicklungen im Bereich sauberer Energie ermöglichen [werden] und nicht nur neue Unternehmen, sondern auch neue Industrien und Verbesserungen der Lebensqualität in fast allen erdenklichen Bereichen hervorbringen [werden].“<sup>3</sup>

Doch neben diesen unglaublichen Chancen birgt der KI-Boom auch Risiken von äußerst besorgniserregendem Ausmaß. Mit KI können wir Systeme schaffen, die sich unserer Kontrolle entziehen und böartigen Akteuren Instrumente an die Hand geben, um Katastrophen unvorstellbaren Ausmaßes auszulösen. KI verändert das technologische Ökosystem; mit KI-gestützter Biotechnologie können wir die Grundbausteine des Lebens manipulieren. KI öffnet die Tür zu automatisierten Cyberangriffen, Kriegen und künstlich erzeugten Pandemien, die ganze Länder verwüsten könnten.<sup>4</sup>

Die technologischen Fortschritte der letzten Jahre haben uns Werkzeuge an die Hand gegeben, die noch vor einem Jahrzehnt unvorstellbar waren. Die Arbeit in der Diakonie bleibt davon nicht unberührt, kann davon nicht unberührt bleiben. Sie wird sich nachhaltig verändern, wie überhaupt das Arbeiten in der Gesundheits- und Sozialwirtschaft sich rasant und grundlegend weiterentwickeln und verändern wird. Was bedeutet diese erneute Revolution in der technologischen Entwicklung für die Akteure der Gesundheits- und Sozialwirtschaft, für die Träger der freien Wohlfahrtspflege, für die Diakonie insbesondere?

Eine weitreichende Antwort scheint derzeit noch nicht möglich, aber erste, sehr vorläufige Antwortversuche möchte ich in diesen Ausführungen geben. Sie gehen von der These Suleymans aus, dass die Welle nicht aufzuhalten ist, dass sie aber nicht zwangsläufig im freien Lauf ein Tsunami werden muss, sondern gelenkt ein heilsamer, menschenfördernder Strom werden kann.

Im Folgenden skizziere ich ausgehend von einer fiktiven Vision des Jahres 2040 Perspektiven auf KI aus diakonischer Sicht, die sich zwischen

---

3 *Ebd.*

4 *Vgl. a.a.O., S. 11–27.*

Heilserwartungen und apokalyptischen Befürchtungen bewegen. Ich werde anschließend die Bedarfe darlegen, die wir als Diakonie an die Künstliche Intelligenz haben, Anwendungsbeispiele und die KI-Leitlinien des EWDE vorstellen, sowie ethische Perspektiven und abschließende Thesen zum künftigen Umgang mit KI präsentieren.

### ***1. KI in diakonischen Unternehmen – zwischen Heilserwartungen und Apokalypse***

Stellen Sie sich vor, wir schreiben das Jahr 2040. Darf ich Ihnen Anne vorstellen? Sie ist Teamleiterin in einem diakonischen Pflegeunternehmen, in dem KI-gestützte Systeme allgegenwärtig sind. Intelligente Sensoren überwachen kontinuierlich die Vitalparameter sowie die emotionalen Zustände der Bewohnerinnen und Bewohner und melden Veränderungen an das Pflegepersonal. Diese Technologie ermöglicht eine ganzheitliche Betreuung, die sowohl körperliche als auch emotionale Bedürfnisse berücksichtigt.

Die Bewohnerinnen und Bewohner haben einen individuellen KI-generierten Pflegeplan, der medizinische Bedürfnisse und persönliche Vorlieben ideal verarbeitet, um das Wohlbefinden zu maximieren. Anne und ihr Team nutzen diese Informationen, um maßgeschneiderte Pflege und Betreuung anzubieten.

KI übernimmt viele administrative Aufgaben, von der Dokumentation bis zur Medikamentenverwaltung, was dem Team mehr Zeit für direkte Pflege ermöglicht.

Anne und ihr Team nutzen KI-Chatbots, die rund um die Uhr verfügbar sind. Diese bieten Echtzeit-Übersetzungen und sind auf den Sprachgebrauch der Teammitglieder trainiert, die überwiegend aus dem Ausland kommen, sodass eine persönliche und reibungslose Kommunikation ermöglicht wird.

Trotz der technologischen Fortschritte bleibt die Menschlichkeit im Mittelpunkt von Annes Arbeit. Sie achtet darauf, dass die ethischen Standards der Diakonie gewahrt bleiben und dass Technologie die menschliche Verbindung unterstützt, nicht ersetzt. Die Pflege hat ein neues Zeitalter erreicht, in dem Herausforderungen wie der Fachkräftemangel durch innovative KI-Lösungen überwunden wurden.

Und ich möchte Ihnen eine weitere Person vorstellen: Das ist Lukas. Ein junger Mann, auch er lebt im Jahr 2040. Lukas ist zunehmend mit KI-Assistenzsystemen aufgewachsen, sie gehören selbstverständlich in sein privates wie berufliches Leben. Seine persönliche KI-Assistenz plant seinen Alltag, seinen Job und seine Kommunikation. Dieses Assistenztool hat ihn bereits durch die Schulzeit begleitet und unterstützt ihn nun in seinen ersten beruflichen Schritten.

Lukas hatte Schwierigkeiten, eine Arbeitsstelle zu finden. Aufgrund seiner genetischen Veranlagung hat er ein erhöhtes Risiko, häufiger und früher als Gleichaltrige zu erkranken. Diese Gesundheitsdaten führten dazu, dass er auf dem regulären Ausbildungsmarkt keine Chance hatte, da die KI-gesteuerte Vorauswahl der Unternehmen ihm kein Bewerbungsgespräch ermöglichte. In letzter Minute fand er eine Anstellung bei einem Träger der Sozialwirtschaft, der bewusst auf KI-gesteuerte Bewerbungsverfahren verzichtet, um Diskriminierung zu vermeiden.

Lukas hat nun eine Stelle in der Drittmittelförderung bekommen. Das gesamte Team arbeitet ausschließlich remote, einen Büroraum oder gar Kolleginnen und Kollegen hat er noch nie außerhalb der Kacheln gesehen. Sowieso schon recht vereinsamt fürchtet er nun um seinen Arbeitsplatz, da auch in der Sozialwirtschaft zunehmend Stellen der KI-Automatisierung zum Opfer fallen. Seine Aufgaben könnten schon bald billiger und effizienter von seinem eigenen Assistenzsystem erledigt werden.

Die fiktiven Charaktere Anne und Lukas aus dem Jahr 2040 symbolisieren die Extreme in der gesellschaftlichen wie diakonischen Diskussion über Künstliche Intelligenz: von Heilserwartungen bis hin zu apokalyptischen Befürchtungen. Doch weder wird KI all unsere Probleme lösen, noch werden posthumane Maschinenwesen uns ersetzen. Diese Szenarien markieren jedoch Grenzen, innerhalb derer wir unsere Zukunft aktiv gestalten können.

Die Suchbewegung in der Entwicklung und Anwendung von KI bewegt sich zwischen Euphorie und Skepsis. Um die Chancen bei einer ausgewogenen Betrachtung der Risiken fruchtbar zu machen, brauchen wir einen nüchternen, pragmatischen Blick auf die realistische Nutzung von KI in der Sozialwirtschaft. Es geht nicht um ein Entweder-oder, sondern um ein Sowohl-als-auch: die Chancen nutzen und die Risiken erkennen. Jetzt ist nicht die Zeit, um in Technikverdrossenheit zu verfallen, oder

die Augen vor der Komplexität von KI zu verschließen. Vielmehr haben wir eine große Aufgabe vor uns: Wir müssen eine Balance entwickeln, in der wir das Gute von KI für die uns anvertrauten Menschen nutzen und zugleich den Prozess aktiv mitgestalten, die Nutzung von KI so einzudämmen, dass katastrophale Folgen verhindert werden.

## ***2. Der diakonische Bedarf an KI***

Warum brauchen wir KI in der Diakonie? Und was genau an KI brauchen wir?

Diese Fragen werden Sie alle unterschiedlich beantworten, je nach Handlungsfeld und konkreter Herausforderung. Übergreifend lässt sich jedoch festhalten: KI kann uns helfen, effizienter zu arbeiten, indem sie administrative Aufgaben automatisiert und uns mehr Zeit für das Wesentliche gibt: die persönliche Betreuung und Unterstützung der Menschen.

Ich möchte drei zentrale Aspekte beleuchten, die unseren gesamt diakonischen Bedarf an KI ausmachen: erforderliches Wissen, erforderliche Technik und eine erforderliche Haltung, um den Wandel gemeinsam zu gestalten.

### ***2.1 Erforderliches Wissen***

Begriffe wie Algorithmen, neuronale Netze, Deep Learning und Large Language Models (LLMs) sind bereits oder werden bald Teil unseres Alltags sein. Auch wenn wir keine IT-Spezialistinnen und Spezialisten werden, müssen wir uns mit einem Basiswissen über KI ausstatten, um die Transformation aktiv mitzugestalten. So wie wir uns in den Sozialgesetzbüchern und politischen Strukturen auskennen, wird auch das Verständnis von KI unverzichtbar werden.

Dieses Wissen umfasst auch die Bewertung von Herausforderungen und ethischen Implikationen, die mit der Implementierung von KI-Tools einhergehen – unter anderem verdeckte Machtstrukturen und kommerzielle Interessen.

Eine große Herausforderung wird die Integration dieses Wissens in die Aus-, Fort- und Weiterbildung sein. Veränderungsprozesse sind oft

schmerzhaft, da sie alte Sicherheiten in Frage stellen. Sorgen und Ängste von Führungskräften und Mitarbeitenden müssen ernst genommen werden: vor Stellenabbau, fehlender Kompetenz und Machtverlust. Ein transparentes und partizipatives Vorgehen bei der Einführung von KI-Tools ist daher unerlässlich.

Zuletzt: Das Wissen über KI verbreitet sich dynamisch und hierarchieübergreifend – dieses Potenzial sollten wir nutzen, indem wir Menschen mit Interesse fördern und in junge Talente investieren.

## ***2.2 Erforderliche Technik***

Neben dem Wissen benötigen wir die erforderliche Technik, um die Chancen von KI in unsere diakonischen Handlungsfelder zu integrieren.

Eine datengeschützte Umgebung für Tools wie ChatGPT ist dafür grundlegende Voraussetzung. Die Auswahl weiterer Tools sollte auf das jeweilige Handlungsfeld abgestimmt sein, worauf ich später noch eingehen werde.

Weiterhin ist auch die Kosteneffizienz entscheidend, da sowohl Technik als auch Schulung kostenintensiv sind und verantwortungsvoll geplant werden müssen. Da wir momentan nicht davon ausgehen können, dass die Kosten der KI-Transformation in nennenswerter Weise in Refinanzierungsstrukturen abgebildet werden, ist es umso wichtiger, dass wir uns intern vernetzen und kooperativ KI-Tools erproben.

## ***2.3 Erforderliche Haltung***

KI ist bereits Realität und wird weiter an Bedeutung gewinnen. Wir sollten diese Entwicklung nicht versuchen zu verhindern, sondern sie proaktiv mitgestalten. Andernfalls wird sich die sogenannte „Schatten-KI“ ausbreiten, indem Mitarbeitende private KI-Tools nutzen.

Vor aller Wissensaneignung oder Implementierung von Tools ist also eine entsprechende Haltung zu KI entscheidend: Wir haben die Chance, die Einführung von KI verantwortungsvoll zu gestalten und Leitlinien zu entwickeln, um die Möglichkeiten, die sie bietet, optimal zu nutzen.

### ***3. Anwendungsbereiche, KI-Leitlinien und KI-Verbundprojekt***

Damit komme ich zum dritten Abschnitt: Anwendungsbereiche, die KI-Leitlinien und das KI-Verbundprojekt der Diakonie Deutschland.

#### ***3.1 Beispielhafte Anwendungsbereiche***

Wo setzen wir in der Diakonie bereits KI ein, und wo eröffnen sich neue Möglichkeiten?

Sie alle haben sicherlich Anwendungsfälle aus Ihrem Arbeitsbereich im Kopf und werden im Verlauf der Tagung viele weitere kennenlernen.

Beispielhaft nenne ich einige diakonische Anwendungsbereiche:

- KI kann in der Pflege und Betreuung eingesetzt werden, um Routinetätigkeiten zu automatisieren, wie z.B. die Medikamentenvergabe oder die Überwachung von Vitalzeichen, wodurch Pflegekräfte entlastet werden. In Frankfurt bei der Agaplesion gAG habe ich von einer beeindruckenden Entwicklung der KI-gesteuerten Dienstplanung erfahren, eine roboterassistierte Operation miterlebt und neueste Entwicklungen einer KI-gesteuerten Strahlentherapie vor Ort kennenlernen dürfen – hier wird sich in Zukunft noch vieles entwickeln. Zum Wohle der uns anvertrauten Menschen!
- In der Sozialarbeit kann KI helfen, Bedarfsanalysen durchzuführen und maßgeschneiderte Unterstützungsangebote in der Vorbereitung auf Beratungsgespräche zu entwickeln. Wir denken über den Einsatz von Chatbots zur Unterstützung in der Beratung nach und nutzen bereits in vielfältigen Kontexten Übersetzungstools.

Ein Beispiel ist die LeavingCare.AI, eine KI-basierte interdisziplinäre Fallberatung, die Fachkräfte der Jugendhilfe im Bereich „Careleaver“ unterstützt, um junge Menschen besser beim Übergang von der Jugendhilfe ins eigenständige Leben zu begleiten. Die App der Social Impact gGmbH befindet sich derzeit in der Pilotphase. Unter dem Motto „Hilfe, um zu helfen“ ermöglicht sie, Anliegen unkompliziert in korrekte Behörden-

kommunikation zu formulieren oder erklärt komplizierte Inhalte der Sozialgesetzbücher in verständliche Sprache.<sup>5</sup>

- KI kann in der Verwaltung eingesetzt werden, um Prozesse zu optimieren und damit zu entlasten und die Effizienz zu steigern. KI unterstützt bei der Informationsaufbereitung, analysiert Daten und wertet sie nach vorgegebenen Kriterien aus, assistiert bei der Texterstellung und formuliert in Behördensprache um.
- KI kann uns ebenso helfen, die Kommunikation und Vernetzung innerhalb unserer Organisationen zu verbessern. Durch den Einsatz von KI-gestützten Kommunikationsplattformen können wir den Austausch von Informationen und die Zusammenarbeit zwischen verschiedenen Abteilungen und Standorten erleichtern.

Die Chancen KI-gestützter Assistenzsysteme sind offensichtlich:

Durch Übersetzungen in Echtzeit werden Sprachbarrieren abgebaut, KI-Tools arbeiten unterstützend rund um die Uhr, sie ermöglichen Zeitersparnis und Entlastung, senken die Fehlerquote gegen null. In der Medizin sind revolutionäre Weiterentwicklungen in Diagnose und Therapie bereits Wirklichkeit.

Doch neben all diesen Möglichkeiten sollten wir auch die Risiken im Blick behalten:

Uns fehlt im weitesten Sinne die Transparenz hinter den Algorithmen. Kommunikation wird mehr und mehr entmenschlicht. Vielen Zielgruppen fehlt die zur Teilhabe erforderliche digitale Kompetenz oder gar der Zugang zu digitalen Endgeräten. KI verstärkt Diskriminierung und Stereotype. Und je nach Nutzung der KI-Tools steht der Schutz unserer Daten auf der Kippe.

---

5 Vgl. <https://leavingcare.ai/>

### **3.2 Leitlinien zum Umgang mit KI im Evangelischen Werk für Diakonie und Entwicklung e.V.**

Um die Chancen von KI verantwortungsbewusst zu nutzen, brauchen wir daher klare Leitlinien. Exemplarisch nenne ich an dieser Stelle die Leitlinien des EWDE.<sup>6</sup>

- Die KI-Leitlinien ermöglichen Orientierung, indem sie regulieren, wie wir in unserer Organisation ethisch und sicher mit KI umgehen.
- Sie sind kein Gesetzestext mit Detail-Regelungen, sondern Prinzipien und Rahmenempfehlungen, die auf verschiedene Anwendungsgebiete übertragbar sind.
- Die Leitlinien sind ein knappes Dokument, das Mitarbeitende ermächtigt, innerhalb eines groben Handlungsrahmens mit KI zu arbeiten. Sie sollen nicht von KI abschrecken!
- Vielmehr sollen die Leitlinien Unsicherheiten nehmen, Mitarbeitende stärken und kompetent machen, indem Gefahrenpotentiale benannt und Empfehlungen gegeben werden.
- Mit den Leitlinien erkennen wir an, dass wir uns in einer Experimentierphase befinden und den Prozess regelmäßig überprüfen und weitergestalten.

Konkrete Elemente der Leitlinien sind unter anderem:

- Grundsätzlich ist unsere Haltung: Mitarbeitende sollen nicht ersetzt, sondern befähigt werden.
- Wir verstehen den Umgang mit KI als einen selbstverantwortlichen und zugleich gemeinsamen Lernprozess.
- Wir verpflichten uns zu einem transparenten und wahrhaftigen Umgang mit KI.
- Gesetzliche Vorgaben zum Datenschutz sind einzuhalten. Es dürfen keine sensiblen Daten über das EWDE, Partner oder Individuen eingegeben werden.

---

6 [Leitlinien zur Nutzung von KI im EWDE März 2024 extern.pdf \(diakonie.de\)](#)

- Wir verpflichten uns zu einem verantwortlichen Umgang hinsichtlich der Risiken von KI. Daher gilt das „Do-no-Harm“-Prinzip, das bedeutet, dass KI-Systeme so eingesetzt werden, dass sie keinen Schaden für Individuen oder die Gesellschaft verursachen. Ebenso gilt das „human-in-the-loop“-Prinzip, womit sichergestellt wird, dass Menschen aktiv in den Entscheidungsprozess von KI-Systemen eingebunden werden und keine vollautomatisierten Entscheidungen getroffen werden. Zuletzt sind wir uns darüber bewusst, dass KI eine mögliche Quelle von Bias, Diskriminierung und Missbrauch ist.

Um von der Leitlinie ins Tun zu kommen, wurde innerhalb des Bundesverbandes der ai-Hub gestartet, der allen Mitarbeitenden im EWDE Zugang zu einer datengeschützten Variante von ChatGPT ermöglicht. Hier ist nicht nur die klassische Textassistentz abrufbar, sondern auch voreingestellte Prompts zur Überprüfung eigener Texte wie beispielsweise ein Bias-Check oder die Übersetzung in einfache Sprache.

### ***3.3 KI-Verbundprojekt***

Neben den KI-Leitlinien und dem ai-Hub haben wir im Sommer dieses Jahres ein KI-Verbundprojekt initiiert, um noch mehr Erfahrungskompetenz zu gewinnen.

Grundlegend gelten dabei die KI-Leitlinien des EWDE, um einen sicheren Umgang zu gewährleisten. Eine breite Einbindung der Helfefelder der Diakonie war in der Projektphase wünschenswert, ebenso wie die Erprobung weiterer KI-Tools neben Chat-GPT.

Unterstützt durch die Firma brandung GmbH sind 290 Teilnehmende aus 130 diakonischen Organisationen als KI-Pioniere unterwegs. Neben Workshops zu KI-Grundlagen beschäftigen sie sich in anwendungszentrierten Gruppen mit den Themen Pflege und Altenhilfe, Öffentlichkeitsarbeit, Fundraising und Fördermittel, Allgemeine Beratung, BWL, Leichte Sprache, Allgemeine Unterstützung sowie Recht und Datenschutz.

Sowohl für den Verband als auch für die beteiligten Organisationen zeigen sich bereits jetzt große Potentiale des Probierraumes: Wissen wird geteilt, der Austausch von Erfahrung wird gefördert und es entstehen konkrete use cases, also spezifische Anwendungsbeispiele für KI in einem Arbeitsbereich, die als Ergebnisse des Projekts in die Arbeitsebene gegeben werden.

## 4. Ethische Perspektiven

Pflegeroboter, KI-generierte Chatbots, roboterassistierte Operationen – diese faszinierenden Möglichkeiten lösen auch Skepsis aus: Wohin wird uns diese Technologie führen?

Vermehrt wird die Forderung nach einem „hippokratischen Eid“ für Informatikerinnen und Informatiker in der Entwicklung von KI laut.<sup>7</sup> Klar ist: Bei aller Faszination über KI dürfen wir die ethischen Implikationen nicht außer Acht lassen. Im weiteren Tagungsprogramm wird das detaillierter betrachtet werden. Heute möchte ich zumindest einige Schlaglichter auf ethische Fragestellungen in Bezug auf KI werfen.

### 4.1 Macht und Verantwortung

Zentral steht die Frage nach Macht und Verantwortung im Zusammenhang mit KI.

Wer entscheidet über die Trainingsdaten eines KI-Systems?

Nach welchen ethischen Standards sollen KI-Systeme agieren?

Ein Beispiel aus dem kirchlichen Bereich zeigt, welche Verantwortung denjenigen zukommt, die KI-Systeme programmieren: Im Rahmen des Reformationsjubiläums wurde von der Evangelischen Kirche in Hessen und Nassau ein Martin-Luther-Avatar entwickelt.<sup>8</sup>

Mit dem Luther-Avatar kann man sich unterhalten, so soll mittels Künstlicher Intelligenz die Botschaft der Reformation auch heute erfahrbar werden, zum Beispiel bei Führungen an historischen Orten oder auch zu religionspädagogischen Zwecken in Schulen.

Eine der Fragen der beiden Entwickler war, mit welchen Daten der Avatar trainiert werden sollte: Mit allen uns vorliegenden Daten, damit es ein historisch korrekter Luther wird, oder sollen nur diejenigen Texte zu Trainingszwecken genutzt werden, in denen sich Luther nicht antisemitisch äußert? Letzteres ist hier der Fall, doch zeigt dieses Beispiel die Tragweite der Verantwortung in der Programmierung von KI-Tools.

---

7 Vgl. u.a. [Der hippokratische Eid – aber digital \(healthcare-digital.de\)](#) oder [Digitalisierung: Der hippokratische Eid für die IT | ZEIT Arbeit](#)

8 Vgl. [Künstliche Intelligenz: Kommunizieren mit Luther-Avatar – EKHN](#)

Macht und Verantwortung – im Raum steht auch die Frage nach der sogenannten „Verantwortungsdiffusion“. Wer trägt die Verantwortung für einen Verkehrsunfall, den ein KI-gesteuertes Fahrzeug verursacht? Technologiehersteller fordern entsprechend „Persönlichkeitsrechte“ für hochentwickelte KI-Systeme, was die Verantwortung diffundieren könnte: Kein Mensch wäre dann dafür verantwortlich, wenn ein KI-System mit Waffen experimentiert oder neue Medikamente herstellt.

Ganz klar ist: Das darf nicht passieren. Wir müssen sicherstellen, dass der Grundsatz „human-in-the-loop“ gewahrt bleibt, zumindest in den Bereichen, die wir diakonisch verantworten. Hier denke ich exemplarisch an die Verantwortung im Kontext von Pflegerobotern oder auch Beratungschabots.<sup>9</sup>

#### **4.2 Datenschutz**

Ein weiteres ethisches Schlaglicht fällt auf den Datenschutz. Hier sind für den diakonischen Bereich insbesondere zwei Perspektiven zu bedenken:

Unternehmensintern müssen Betriebsgeheimnisse geschützt werden. Daher ist es bedeutsam, dass wir eigene, datengeschützte Varianten der KI-Systeme wie ChatGPT nutzen.

Eine zweite Perspektive richtet sich auf die personenbezogenen Daten, sowohl der Mitarbeiterinnen und Mitarbeiter als auch der Klientinnen und Klienten. Im Fokus stehen hier insbesondere deren Gesundheitsdaten, die im Umgang mit KI-Tools geschützt werden müssen.

Training und Betrieb von KI-Systemen führt zu hohen Risiken im Bereich des Datenmissbrauchs. Um die Datenverwendung besser kontrollieren zu können, ist im August dieses Jahres die europäische KI-Verord-

---

9 *In der Diakonie stehen wir in diesem Zusammenhang vor einem weiteren Spannungsfeld: Uns fehlen sowohl die finanziellen als auch die personellen Ressourcen, um KI-Systeme nach unseren ethischen Maßstäben zu trainieren. Daher sind wir auf externe Spezialist:innen angewiesen, die diese Tools für uns programmieren. In der Diskussion um Macht und Verantwortung tritt zudem ein weiteres ethisches Problem zutage: Überwiegend Menschen aus Drittländern trainieren gegen Billiglohn die großen KI-Systeme, indem sie Daten auswerten. Dabei sind sie teilweise stundenlang Gewaltdarstellungen ausgesetzt, die sie zu KI-Trainingszwecken identifizieren und ausfiltern müssen. Vgl. [KI: Wie Klickarbeiter in Kenia ausgebeutet werden | tagesschau.de](#)*

nung, der „AI Act“, in Kraft getreten. Die europäischen Mitgliedsstaaten müssen nun Behörden benennen, die die Umsetzung der Regeln national beaufsichtigen. Die europäische KI-Verordnung stuft mehrere KI-Anwendungen als hochriskant ein und formuliert strenge Anforderungen: etwa bei der Bewerberauswahl, in der Justiz, bei Grenzkontrollen oder im Bildungswesen.<sup>10</sup>

In unserem Kontext ist es besonders interessant, dass der europäische „AI Act“ nicht nur die Verantwortung der Anbieter von KI-Systemen hervorhebt, sondern auch die der Betreiber. Das betrifft all jene, die KI-Systeme eigenverantwortlich im Rahmen ihrer beruflichen Tätigkeit nutzen. Ab Februar 2025 tritt die Regelung in Kraft, dass alle Mitarbeiterinnen und Mitarbeiter, die mit KI-Systemen arbeiten, über ein „ausreichendes Maß an KI-Kompetenz“<sup>11</sup> verfügen müssen. Dies macht deutlich: Als Führungskräfte sind wir verpflichtet, unseren Mitarbeiterinnen und Mitarbeitern entsprechende Schulungen anzubieten, damit sie sicher und kompetent mit KI umgehen können.

### ***4.3 KI und synthetische Biologie***

Ein Blick in die Zukunft zeigt, dass die Verbindung von KI und synthetischer Biologie zunehmend an Bedeutung gewinnen wird. Die Gefahr besteht darin, dass KI zur Gestaltung oder Veränderung biologischer Systeme eingesetzt werden kann, was zu Missbrauch und unvorhersehbaren Folgen führen könnte, insbesondere wenn Sicherheits- und Regulierungsmaßnahmen unzureichend sind.

Ein Beispiel hierfür ist die Möglichkeit, dass KI-gestützte Algorithmen neue, synthetische Viren entwerfen. Solche Viren könnten unbeabsichtigt oder absichtlich freigesetzt werden und erhebliche gesundheitliche und ökologische Schäden verursachen – die Coronapandemie hat uns bereits einen Eindruck davon vermittelt, was das bedeuten könnte.

Ein weiteres Beispiel ist die Kombination von KI mit der Gen-Optimierung durch die sogenannte „Genschere“ (CRISPR-Cas9). Diese Technologie ermöglicht es bereits heute, DNA präzise zu schneiden und zu verändern. Mit Unterstützung von KI kann dieses Verfahren erheblich

---

10 [Vgl. \*Wer kontrolliert die Künstliche Intelligenz?\* | \*tagesschau.de\*](#)

11 [Art. 4 KI-VO – KI-Kompetenz – KI-Verordnung \(ai-act-law.eu\)](#)

schneller, exakter und kostengünstiger durchgeführt werden. Die Fähigkeit, Gene präzise zu verändern, könnte jedoch auf dem globalen Biotechnologiemarkt ohne angemessene Regulierung für unethische Zwecke genutzt werden, wie etwa bei der Schaffung von Designerbabys.<sup>12</sup>

#### **4.4 KI verändert Sozialität**

„Alles wirkliche Leben ist Begegnung“, sagte der jüdische Religionsphilosoph Martin Buber. Die Diskussion um KI ist eng mit der Frage verbunden, wo Menschen künftig durch KI ersetzt werden. Ganz klar ist, dass die Bindung von Mensch zu Mensch durch den Einsatz von Sozialrobotern und Chatbots auf verschiedenen Ebenen gelockert werden wird. Wir stehen dadurch vor der Herausforderung, wie wir menschliche Bindung aufrechterhalten und in welchen Bereichen wir keine Lockerung der menschlichen Bindung zulassen oder gar fördern wollen.

Dies spitzt sich in Zeiten des Fachkräftemangels in besonderer Weise zu. Künstliche Intelligenz kann den Fachkräftemangel abmildern, indem sie Routinetätigkeiten übernimmt und so das pflegerische Personal unterstützt. Dennoch ist es wichtig zu betonen, dass der Mangel an Fachkräften nicht unverhältnismäßig durch den Einsatz von KI und Robotik auf Kosten der menschlichen Pflege ausgeglichen werden darf.

#### **4.5 KI und Energiebilanz**

Zuletzt stellt sich die Frage nach der Energiebilanz von KI. Das Training und der Betrieb großer KI-Modelle erfordern erhebliche Energiemengen. Darüber hinaus wird auch viel Wasser benötigt, insbesondere zur Kühlung der Rechenzentren. Laut einer Studie der University of California hat allein das Training von ChatGPT-3 etwa 5,4 Millionen Liter Wasser verbraucht.<sup>13</sup>

Um den enormen Energiebedarf zu decken und gleichzeitig die Klimaziele zu erreichen, setzen Tech-Giganten wie Google und Microsoft zunehmend auf Nuklearenergie, was von einigen bereits als deren

---

12 Vgl. *Mustafa Suleyman: The coming wave. Künstliche Intelligenz, Macht und das größte Dilemma des 21. Jahrhunderts*, München 2024, S. 96–110.

13 Vgl. [Warum KI viel Wasser und Strom braucht | tagesschau.de](https://www.tagesschau.de/wirtschaft/kunstliche-intelligenz/wasser-und-strom-101.html)

Renaissance bezeichnet wird. Allerdings bleibt die Entsorgung des Atommülls ein weltweit ungelöstes Problem.

Es ist wichtig, achtsam zu sein, um unsere Ziele nicht gegeneinander auszuspielen und die Bestrebungen nach Nachhaltigkeit, auch in der Diakonie, nicht zu gefährden.

## *5. Sechs Thesen für die zukünftige Nutzung von KI*

Ausgehend von den fiktiven Personen Anne und Lukas aus dem Jahr 2040 habe ich Ihnen in diesem einleitenden Vortrag im Schnelldurchlauf einige wenige relevante Aspekte von KI für die Diakonie dargelegt.

Insgesamt ist es hochspannend, dass wir eine Suchbewegung in allen Bereichen unserer Gesellschaft erleben. Die Entwicklung neuer KI-Tools verläuft dynamisch, ebenso bilden sich in der diakonischen Verbandslandschaft in unterschiedlichen Geschwindigkeiten Transfers dieser neuen Möglichkeiten ab. Das Thema der KI ist komplex. Wir sollten uns davon aber nicht lähmen lassen, sondern genau hinschauen und strategisch überlegen, welche Chancen der KI wir unter Berücksichtigung der Risiken für unsere Zwecke nutzen können. Die Losung für den Kirchentag in Hannover 2025 möge uns dabei leiten: „mutig – stark – beherzt“!

Zum Abschluss und zugleich als Einladung zur Diskussion stelle ich Ihnen sechs Thesen für einen zukünftigen Umgang mit KI vor:

- 1. Der Mensch im Mittelpunkt:** KI soll den Menschen unterstützen, nicht ersetzen. Menschliche Begegnungen müssen bewahrt werden. Das gilt insbesondere für den Bereich der Pflege. Bei aller Möglichkeit der KI-Assistenz darf die Pflege nicht entpersonalisiert werden. Wir müssen menschliche Nähe bewahren, wo sie erforderlich ist. Zugleich gilt es, die revolutionären Veränderungen in der Diagnostik, Therapie und Beratung durch KI zum Wohl der Patientinnen und Patienten zu nutzen! Hier liegen fast unvorstellbare Potentiale.
- 2. Transparenz und Verantwortung:** Der Einsatz von KI muss transparent gestaltet sein, um das Vertrauen der Mitarbeitenden sowie der Klientinnen und Klienten zu bewahren. Nach dem Prinzip „Human in the loop“ muss in allen diakonischen Abläufen die Letztverantwortung in menschlicher Hand bleiben. Dies gilt insbesondere für sensible Entscheidungen wie beispielsweise in der Personalauswahl.

3. **Datenschutz und Sicherheit:** Der Schutz sensibler Daten der Organisation sowie aller beteiligten Individuen muss oberste Priorität haben, auch, wenn dies im Alltag zu Zielkonflikten mit der Einführung von KI-Tools führen wird.
4. **Kontinuierliche Weiterbildung:** Wir sollten darauf vorbereitet sein, dass unsere Projektpartner ebenso wie Behörden und Stiftungen künftig auf den Einsatz von KI setzen werden. Daher müssen Mitarbeitende im Umgang mit KI geschult werden, damit wir Prozesse mitgestalten können und auch bei der Vergabe von Drittmitteln weiterhin berücksichtigt werden. Die Implementierung von KI-Tools wird in vielerlei Hinsicht ein Kraftakt für einzelne diakonische Unternehmen sein. Daher ist es wichtig, dass wir uns vernetzen und gemeinsam Erfahrungen und Ressourcen teilen.
5. **Ethische Verantwortung:** Wir müssen sicherstellen, dass die Anwendung von KI-Tools im Rahmen der Diakonie unseren Werten entspricht. Daher ist es erforderlich, dass wir ethische Leitlinien im Umgang mit KI entwickeln. Das erfordert eine aktive Grundhaltung: Wir haben die Chance, die Vorteile von KI für das Wohl der uns anvertrauten Menschen zu nutzen! Zugleich stehen wir in der Verantwortung, diese Nutzung so zu steuern, dass Schaden vermieden wird.
6. **KI entscheidet über Zukunft der Diakonie:** Ohne eine strategische Investition in KI-Technologien riskieren wir, im Zeitalter der Künstlichen Intelligenz von anderen sozialen wie wirtschaftlichen Akteuren nicht wahrgenommen oder gar abgehängt zu werden. Daher ist es von entscheidender Bedeutung, dass wir finanzielle Ressourcen freisetzen und neue Mittel generieren, um in KI zu investieren. Nur so können wir sicherstellen, dass die Diakonie auch in Zukunft eine bedeutende Rolle im Sozialwesen einnimmt.

*Theresa Züger*

## WIE KANN KI IM SINNE DES GEMEINWOHLS INGESETZT WERDEN?

### *1. Zum Begriff „Gemeinwohl“*

Um die leitende Frage meines Beitrags zu beantworten, müssen wir zunächst einmal Gemeinwohl definieren. Der Begriff wird häufig gebraucht, aber ihn zu definieren, ist gar nicht so leicht.

In der Arbeitsgruppe „Public Interest AI“ haben wir deshalb erst einmal die politische Theorie befragt, welche Gemeinwohlbegriffe sich dort finden. Dabei hat sich die Definition von Barry Bozeman als hilfreich erwiesen. Er hat vor allem unter Bezug auf John Dewey gearbeitet. Seiner Ansicht nach handelt es sich beim Gemeinwohl um diejenigen Resultate, die auf lange Sicht gesehen am besten dem Überleben und Wohlergehen eines sozialen Kollektivs, verstanden als Öffentlichkeit, dienen.<sup>1</sup>

Aus dieser Definition lassen sich die folgenden Punkte ableiten: Zunächst impliziert die Definition, dass es keine substanzielle und universell gültige Definition von Gemeinwohl gibt. Was Gemeinwohl ist, muss konkret an den Problemen einer faktischen Gesellschaft entlang ausgehandelt werden – und zwar für jedes auftretende Problem neu.

Zwar sind in vielen demokratischen Verfassungen Gemeinwohlprinzipien verankert und teils sehr stark mit unserem demokratischen Selbstverständnis verknüpft. Aber, was Gemeinwohl konkret bedeutet, das muss stets neu ausgehandelt werden. Insofern lässt sich Gemeinwohl als ein prozeduraler Begriff verstehen. Dabei kommen jedoch außerdem gemeinsame Werte in Form eines geteilten Leitbildes von Gemeinwohl zum Tragen, weshalb Gemeinwohl von Bozeman auch von einem prozeduralistischen und idealistischen Ansatz spricht. Damit die Aushandlung gelingt, müssen Bürgerinnen und Bürger, ohne individuelle Interessen einzubringen und ohne auf private Gewinnmaximierung aus zu sein, über die Frage: „Was ist denn gut für uns als Gesellschaft?“ deliberieren.

---

<sup>1</sup> Barry Bozeman: *Public values and public interest counterbalancing economic individualism*, Georgetown 2007.

Nur auf diesem Weg kann eine an Gemeinwohl orientierte Lösung entstehen. Dafür benötigen wir bestehende Ansatzpunkte für einen Prozess der Aushandlung, wie ihn bspw. unsere Verfassungen liefern. Um eine inklusive Aushandlung von Gemeinwohlfragen zu ermöglichen, braucht es stets partizipative Einstiegspunkte für die Deliberation und die Verhandlung aller relevanten Interessen. Daher rührt auch die enge Verbindung, welche Gemeinwohl mit den Prinzipien der Gleichheit und Gleichberechtigung aufweist, die sowohl in der rechtlichen Aushandlung von Gemeinwohlfragen als auch in der Rechtsphilosophie zum Tragen kommt.<sup>2</sup> Beide stellen Voraussetzungen dar, ohne die von Gemeinwohl gar nicht gesprochen werden kann.

Kritisch lässt sich daraus ableiten, dass eine gemeinwohlorientierte Perspektive tendenziell einer Orientierung an Profitmaximierung durch Einführung neuer Anwendungen entgegensteht. Denn sich am Gemeinwohl zu orientieren, heißt auch, zu fragen, wer denn neue Anwendungen eigentlich besitzen sollte und nicht, wer sie faktisch besitzt.

## *2. Das Vorgehen der Forschungsgruppe*

Unser Hauptaugenmerk haben wir in unserer Forschungsgruppe nicht auf die „Mainstream-KI-Industrie“ gelegt. Vielmehr haben wir Projekte in den Mittelpunkt gestellt, die von vornherein auf einen gemeinwohlorientierten Ansatz gesetzt haben. Von diesen versuchen wir im Sinne von Best Practice zu lernen. Dabei fragen wir uns bspw.: Was machen diese Projekte anders als die Mainstreamindustrie? Wie entwickeln sie KI und können wir davon Ideen ableiten, die auch für die Mainstreamindustrie relevant sind?

Um diese Fragen zu beantworten, erheben wir Daten dieser Projekte und sammeln sie in einer Datenbank. Ich muss aber gleich hinzusetzen, dass das noch eine kleine Nische darstellt angesichts der Masse an profitgetriebenen KI-Projekten. Sowohl Informationen zu Projekten als auch Ergebnisse unserer theoretischen Arbeit zu gemeinwohlorientierter KI wie die von uns abgeleiteten Prinzipien und Vorgehensweisen findet man auf der Seite [publicinterest.ai](https://publicinterest.ai) zusammengestellt.

---

2 Mike Feintuck: *'The Public Interest' in Regulation*, Oxford 2004.

Um nicht nur theoretisch von Prinzipien zu sprechen, haben wir auch selbst KI-Anwendungen entwickelt, um zu sehen, wo in der faktischen Umsetzung die Prinzipien dann auch zum Tragen kommen und welchen Hürden die tatsächliche Entwicklung von gemeinwohlorientierten KI-Anwendungen man begegnen kann.

Mein Beitrag widmet sich sowohl diesen Prinzipien als auch der Frage nach dem Wesen von KI und schließlich einzelnen Anwendungen, die wir gemeinsam analysieren.

### ***3. Prinzipien gemeinwohlorientierter KI***

Ich beginne mit den Prinzipien: Das erste zentrale Prinzip nennen wir „Justification“. Unter diesem Begriff befragen wir KI daraufhin, ob das zugrundeliegende Problem tatsächlich mittels KI bearbeitet werden sollte, also ob es eine Rechtfertigung für den KI-basierten Lösungsansatz für ein bestimmtes soziales Problem gibt. Im Raum steht damit nämlich immer auch die Option, ob nicht andere, technisch robustere oder einfachere Lösungen ebenfalls denkbar wären. Denn nicht immer ist transparent, warum ein Problem eigentlich gerade mittels KI bearbeitet werden sollte. Angesichts der obigen Definition von Gemeinwohl sollte der Einsatz von KI angesichts verschiedenster Probleme abgewogen werden. KI-Systeme sind oft aufwändiger, ressourcenintensiver und intransparenter als andere technische Systeme.

Ein zweiter Punkt wird mit „Equity“ betitelt. Hier geht es um die Gleichberechtigung von Menschen, die im besten Fall durch die KI-Anwendung gestärkt werden sollten. Unter dieser Überschrift wäre zu fragen, ob der Einsatz von KI die Gleichberechtigung fördert oder Ungleichbehandlung, z.B. durch Diskriminierung und Biases, verstärkt. Positiv würde hier bspw. ein Projekt bewertet, das die Barrierefreiheit erhöht. Zwar lassen sich vielleicht nicht für alle gesellschaftlichen Gruppen in gleichem Maße positive Effekte erzielen, aber als gesamtgesellschaftliche Zielrichtung ist Equity wichtig.

Drittens sollten KI-Projekte auf einen partizipativen Designprozess Wert legen und Prozesse der Deliberation zulassen. Sie sollten Stakeholder, Menschen, die von der Technologie betroffen sein werden, in den Gestaltungsprozess involvieren. Und zwar nicht erst am Ende, um die dann fertige Technologie zu testen, sondern bereits im Entwicklungs-

prozess. Die Betroffenen sollten also mitreden können, wenn es um die grundlegende Ausrichtung der Technologie geht. Partizipatives Design kann dabei sehr unterschiedliche Formen und je nach Projekt auch eine unterschiedliche Intensität annehmen. Partizipation ist sowohl finanziell als auch methodisch eine zusätzliche Hürde für eine Projektentwicklung, doch zeigt sich in der Praxis immer wieder, dass Projekte auch und gerade daran scheitern können, dass sie zu wenige relevante Stimmen inkludiert haben und zu intransparent vorgegangen sind.

Viertens sollten technische Standards und Sicherheitsplanken etabliert sein. Systeme müssen gegen Missbrauch, Hacking, Datenverlust und anderes geschützt werden. Ähnliches gilt für den Fall, dass Systeme nicht den angepriesenen Effekt haben und Erwartungen enttäuschen.

Fünftens sollten die Systeme einer kritischen Evaluierung zugänglich sein. Wissenschaftlerinnen und Wissenschaftler oder andere Prüfinstanzen sollten überprüfen können, wie ein System im Detail funktioniert und ob es das tut, was es verspricht, ohne dabei negative Nebeneffekte zu haben. Eine Voraussetzung dafür ist bspw. die Verfügbarkeit von Daten. Bei großen Modellen weiß beispielsweise niemand genau, mit welchen Daten diese trainiert wurden. Und gerade dort stellen sich auch große Fragen zum Urheberrecht, da viele Daten, die über Scraping im Internet zusammengestellt und für KI-Modelle genutzt werden, dem Urheberrecht unterliegen. An dieser Stelle bedarf es der Transparenz – auch über die Datenquellen von KI-Modellen, damit die Grenzen von Systemen analysiert werden können.

Sechstens muss auch Nachhaltigkeit als Maßstab an KI-Systeme angelegt werden. Zum einen natürlich im Sinne ökologischer Nachhaltigkeit. Teilweise haben KI-Systeme im Training und auch durch ihre Nutzung (die sogenannte Inferenz) einen hohen und vor allem steigenden Energie- und auch Ressourcenverbrauch, z.B. an Wasser und Rohstoffen.<sup>3</sup> Zum anderen geht es bei nachhaltiger KI-Entwicklung auch um Nachnutzbarkeit. Das zielt darauf ab, dass auch weitere Akteure später noch verstehen, wie ein System funktioniert und wie es weiterentwickelt werden kann. Insofern geht es darum, die Systeme wiederverwertbar zu machen. Das ist

---

3 *Hannah Smith/Chris Adams (Green Web Foundation): Thinking about using AI? Here's what you can and (probably) can't change about its environmental impact. <https://www.thegreenwebfoundation.org/publications/report-ai-environmental-impact/>.*

im Moment kaum Standard. Sehr viele Projekte werden von öffentlicher Seite für eine kurze Zeit gefördert und mit dem Zeitpunkt der Beendigung der Förderung verschwinden die Projekte wieder. Selbst die ehemaligen Ansprechpartner lassen sich oftmals nicht mehr ausmachen.



Abbildung 1: Prinzipien für gemeinwohlorientierte KI, wie auf [publicinterest.ai](https://publicinterest.ai)

#### 4. Was ist und wie funktioniert KI?

Was ist also KI genau? Künstliche Intelligenz ist ein komplexer Sammelbegriff, der nur schwer eindeutig definierbar ist. Begriffsgeschichtlich lässt sich sagen: Der Begriff „Künstliche Intelligenz“ wurde Mitte der 1950er Jahre von Wissenschaftlern zum ersten Mal genutzt (McCarthy). Er wurde bewusst gewählt, um Fördermittel einzuwerben. Andere begriffliche Vorschläge waren akkurater, man könnte sagen, wissenschaftlicher. Doch damit vielleicht auch deutlich weniger spannend. Deshalb hat man sich entschieden, von KI zu sprechen. Gleichzeitig war vielen Wissenschaftler\*innen schon damals klar, dass der Begriff auch problematisch ist, weil er dazu verführt, technische Systeme zu vermenschlichen, also fachsprachlich zu anthropomorphisieren. Hier spielt auch der alte Menschheitstraum hinein, eine Maschine zu entwickeln, die dem Menschen gleicht. Er steht häufig im Hintergrund, wenn KI-Technologien überschätzt werden und Projektionsflächen von Träumen als auch Dystopien werden.

Andererseits steht KI für ein Teilgebiet der Informatik, das sich damit beschäftigt, menschliche Intelligenz technisch nachzubilden. KI-Forschung existiert in der Informatik seit vielen Jahrzehnten und verstand unter KI-jeweils unterschiedliche Technologien. Seit den 2010er Jahren versteht man unter KI hauptsächlich sog. „Machine Learning“ und „Deep Learning“. Beides basiert auf dem Einsatz sog. neuronaler Netze, die im Grunde sehr große technische Rechenmaschinen sind. Daten werden dabei nacheinander von verschiedenen Knotenpunkten (Neuronen), die wie ein Netz miteinander verbunden sind, mathematisch verarbeitet. Neuronale Netze können durch Trainingsdaten darauf trainiert werden,

ein Objekt oder ein Muster zu erkennen, wie zum Beispiel Gesichter, oder ein wahrscheinliches nächstes Wort in einer menschlichen Sprache. Dabei kommen stochastische Verfahren zum Einsatz. Es geht also um Wahrscheinlichkeitsprognosen, was bedeutet, das KI-Systeme dieser Art keine regelbasierten Antworten geben, sondern eine Wahrscheinlichkeit bestimmen, mit der jeweils mögliche Ergebnisse zutreffen könnten – zum Beispiel, dass eine Anordnung von Pixeln auf einem Bild mit einer Wahrscheinlichkeit von 90% ein menschliches Gesicht darstellt.

Wie entsteht eine KI-Anwendung? Ein sogenannter KI-Lifecycle sieht meist folgendermaßen aus.

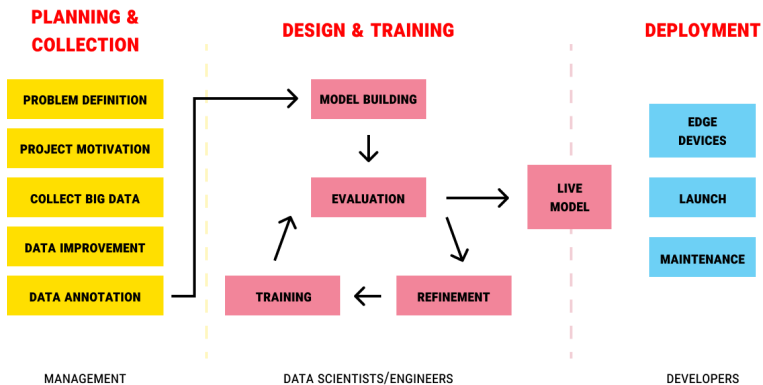


Abbildung 2: KI-Lifecycle, Quelle: <https://labeleyourdata.com/articles/lifecycle-of-an-ai-project-stages-breakdown>

Er beginnt links mit der Identifikation eines Problems, für das ein KI-System als Lösung entwickelt werden soll. Im nächsten Schritt geht es darum, Daten zu akquirieren. Sie stellen die Grundlage dafür dar, ein KI-Modell zu trainieren. Zumeist liegen diese Daten nicht in einer optimalen Form vor, sondern müssen erst aufbereitet werden. In diesen Schritt der Datenaufbereitung fließen meist viel Zeit und Ressourcen. Die Daten müssen möglicherweise maschinell lesbar gemacht werden, kategorisiert und bereinigt werden. Im Regelfall übernehmen diesen Schritt reale Menschen, die Daten bearbeiten. Im nächsten Schritt sind es häufig auch Menschen, die die Annotation von Daten übernehmen, z.B. Objekte kategorisieren oder die Qualität eines Outputs bewerten. Wenn sie ein Capture-Verfahren kennen, das manchmal im Internet zu lösen ist, bevor wir eine Seite ansteuern können, können Sie sich

vorstellen, wie in etwa eine Annotationssaufgabe aussehen kann. Erst durch das Training von Modellen durch die menschlichen Annotationen können viele KI-Systeme die gewünschten Ergebnisse erzielen. Für viele Anwendungen werden heute bereits vortrainierte Modelle genutzt, die dann ggf. nur „gefinetuned“ werden, also für eine spezifische Aufgabe nachtrainiert werden – was weniger Zeit und Ressourcen als die Neuentwicklung erfordert. In der Entwicklung eines KI-Modells braucht es zumeist mehrere Iterationen des Testens und Anpassens der Leistungsfähigkeit eines Modells, bevor dieses zur Nutzung implementiert wird. Auch nach vielen solchen Schleifen können KI-Modelle, wenn ihnen neue Daten oder Aufgaben begegnen, immer noch scheitern, also Fehler produzieren oder eine wesentlich schlechtere Performance an den Tag legen als im Training. Dies erfordert dann eventuell den Trainingsprozess erneut zu verbessern. Erst wenn eine ausreichende Performance erreicht wird, sollten Systeme implementiert werden (siehe Abbildung 2), was jedoch zumeist weiterhin eine Instandhaltung erfordert.

#### 4.1 Neuronale Netzwerke

Das Kernstück einer KI stellt, wie erwähnt, ein neuronales Netz dar. Der Begriff ist inspiriert durch die Neuronen des menschlichen Gehirns, wobei zwischen diesem und technischen neuronalen Netzen weiterhin gravierende Unterschiede existieren. Die Begriffsähnlichkeit sollte nicht dazu verleiten, zu übersehen, dass es sich bei neuronalen Netzen um mathematische Modelle handelt, nicht um materielle Neuronen. Beispielhaft erkläre ich folgend den Aufbau und die Funktionsweise eines neuronalen Netzes anhand der Problemstellung „Gesichtserkennung“. Stellen Sie sich vor, Sie haben ganz viele Bilder und möchten ein System trainieren, das in der Lage ist, zu identifizieren, ob ein Bild ein Gesicht zeigt oder nicht.

Auf der Seite der Ausgabeebene (siehe Abbildung 3, hier in rot) existieren damit zwei Optionen: Gesicht/kein Gesicht. Das neuronale Netz gibt auf dieser Seite einen Wahrscheinlichkeitswert aus: Wie wahrscheinlich es ist, dass auf diesem Bild ein Gesicht zu finden ist – oder eben nicht? Zu diesem Ergebnis kommt es auf dem Weg eines sehr komplexen statistischen Verfahrens. Zwischen dem Input-Layer und dem Output-Layer können viele Zwischenebenen (Hidden Layer) zur Verarbeitung geschaltet sein. Was geschieht nun zwischen Input und Output? Das Bild wird in Zahlen zerlegt, beispielsweise in Pixel und Graustufen-Werte. Diese Werte stellen dann die Eingabe für das neuronale Netz dar. An

jedem Punkt der Layer erfolgt jetzt eine mathematische Berechnung. Die einzelnen Punkte werden in der Bedeutung ihres Ergebnisses gewichtet und geben das Ergebnis an weitere Ebenenpunkte weiter. Die einzelnen Neuronen operieren häufig mit einem Schwellenwert, der überschritten sein muss, damit eine Weitergabe erfolgt. Jede einzelne Berechnung repräsentiert die Erfassung bestimmter Muster innerhalb der Datenstruktur, die final zur Wahrscheinlichkeitsrechnung für die Optionen Gesicht/kein Gesicht ausgewertet werden. Es liegt nahe, hier die vermenschlichende Sprechweise von „entscheiden“ zu verwenden. Das System „entscheide“, hier sei ein Gesicht zu sehen. Doch das ist nicht der Fall, es handelt sich um statistische Berechnungen. Am Ende dieser Berechnungsschleifen steht ein Ergebnis auf der Ausgabeseite.

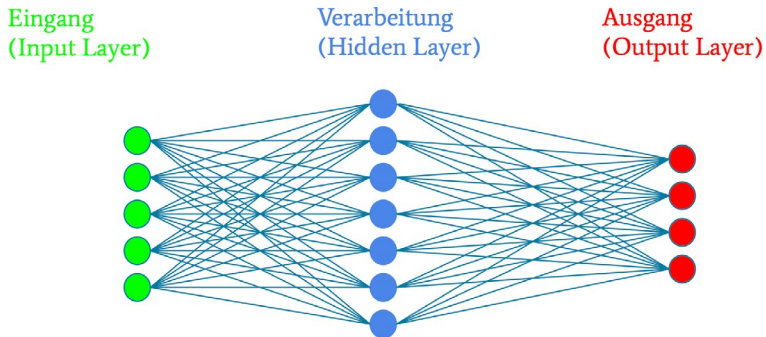


Abbildung 3: Neuronale Netze

Ein solches System lässt sich folgendermaßen trainieren. Sie geben dem System für viele einzelne Bilder im Training die korrekte Antwort, z.B. „Gesicht“, wenn ein solches zu erkennen ist, und das System kalibriert durch eine Rückrechnung die Gewichtungen der Knoten, die Parameter der statistischen Berechnung so, dass die Performance des Systems bei der Erkennung jeweils ein bisschen besser wird. Auch diesen Prozess der Rückrechnung wiederholen sie unzählige Male und justieren damit das neuronale Netz in sehr vielen Wiederholungen dazu, Gesichter (in ihren Unterschieden) mit einer immer höheren Treffsicherheit zu erkennen. Dabei werden durch das System Muster identifiziert, die die Wahrscheinlichkeit erhöhen, dass auf dem Bild ein Gesicht zu finden ist. Diese Muster entsprechen jedoch nicht zwingend denjenigen, die wir Menschen verwenden, um dieselbe Operation auszuführen (wie z.B. „hat eine Nase“ oder „hat zwei Augen“).

## 4.2 NLP – Natural Language Processing

Gehen wir einen Schritt weiter. Beim sog. Natural Language Processing (NLP) tritt noch ein weiterer Schritt hinzu. Es handelt sich dabei zunächst einmal um ein probabilistisches statistisches Modell, das die Wahrscheinlichkeit des Vorkommens einer bestimmten Wortfolge in einem Satz auf Grundlage der vorherigen Wörter bestimmt. Es hilft vorherzusagen, welches Wort mit höherer Wahrscheinlichkeit als nächstes im Satz vorkommt. Es vollzieht also eine grundlegend andere Operation als wir Menschen, wenn wir „Sprache produzieren“. Die Grundlage für NLP bilden ebenfalls neuronale Netze. Aber diese unterscheiden sich bei NLP von den oben vorggeführten. Was sie unterscheidet, wird „Embeddings“, also Einbettungen, genannt.

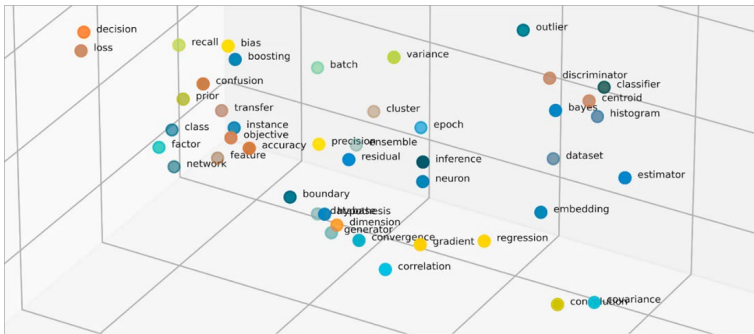


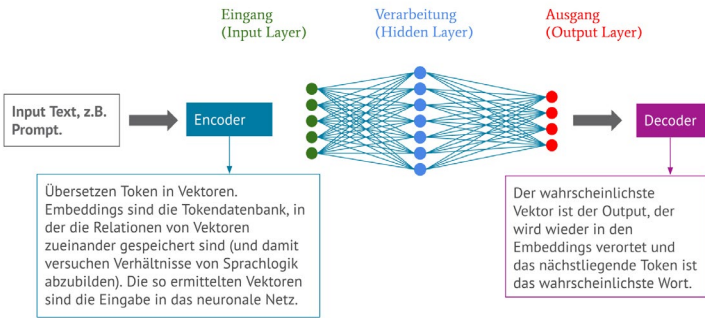
Abbildung 4: Embeddings, Quelle: <https://aws.amazon.com/de/what-is/embeddings-in-machine-learning/>

Metaphorisch kann man sich Einbettungen wie ein mehrdimensionales Wörterbuch vorstellen. In diesem Wörterbuch werden Worte in sog. Tokens zerlegt, also in kleine Wortteile und Schnipsel. Diese Tokens wiederum werden in mathematische Vektoren übersetzt. Dadurch generieren sie einen mehrdimensionalen Vektorraum, in dem jeder Sprachschnipsel in Verhältnisse zu anderen Sprachschnipseln gesetzt wird. Diese mathematische Vektorlogik entspricht nicht der morphologischen Logik unserer Sprache. Vielmehr stellt er eine mathematische Abbildung, ein Modell zur Erklärung menschlicher Sprachprozesse dar. Diese Systeme können also nicht semantische Kontexte verstehen, sondern sie können berechnen, welche Distanz ein bestimmtes Objekt zu einem anderen Objekt in einem mehrdimensionalen Vektorraum besitzt. Diese Systeme liefern also Wahr-

scheinlichkeitswerte für Objektdistanzen in Vektorräumen. Das gleichzusetzen mit dem, was geschieht, wenn Menschen im emphatischen Sinn sprachliche Äußerungen verstehen, wäre ein Missverständnis.

Wenn jemand nun einen Prompt eingibt, also eine sprachliche Eingabe in einem Interface eines KI-Systems tätigt, dann wird diese über einen sog. Encoder in Tokens zerlegt und diese Tokens wiederum werden in Vektoren umgesetzt. Damit repräsentiert das System die sprachliche Eingabe in das System des mathematischen Vektorraums. Die Embeddings speisen diese Token-Datenbank in das neuronale System ein. Auf diese Weise kann ein solches System sinnvolle Tokenkontexte reproduzieren. Im Grunde genommen geschieht dann etwas Vergleichbares wie bei Bildern: es laufen statistische Berechnungen, die Wahrscheinlichkeitswerte ausgeben. In diesem Fall besteht der Output im wahrscheinlichen nächsten Wort. Also zunächst einmal in einem Vektor, dann einem Token und dann einem Wort, solchermaßen muss der Output erst wieder aus dem Mathematischen ins Sprachliche zurückübersetzt werden.

#### NLP und Neuronale Netzwerke



12

AI & SOCIETY LAB AT ALEXANDER VON HUMBOLDT INSTITUTE FOR INTERNET AND SOCIETY

Abbildung 5: Erläuterung zur Funktionalität von LLMs

Grundlegend für diesen Vorgang war die Entwicklung sog. Transformer-Architekturen rund um das Jahr 2017. Diese besitzen die Fähigkeit, vorher bereits bearbeitete Tokens und vorhergehende Wahrscheinlichkeitsberechnungen in die Berechnung des aktuellen Werts mit einzu beziehen. Die damit erfüllte Funktion wird „attention layer“ genannt. Zwar „merkt“ sich diese Ebene nicht tatsächlich etwas. Aber sie speichert vorheriges Sprachgeschehen mathematisch ab und kann so ihre Berechnungen verbessern. Im Grunde handelt es sich dabei einfach um eine

Vergrößerung des zur Verfügung stehenden Datensatzes zur besseren Vorhersage des nächsten Wortes.

### 5. Gemeinwohlorientierte KI: Aufgaben, Einsatzgebiete, Probleme

Nun zur zweiten Frage: Was lässt sich mit den bis hierher skizzierten technischen Systemen an gemeinwohlorientierten Aufgaben leisten?

Unser Datensatz umfasst ca. 250 Projekte. Diese haben wir anhand der SDGs, Sustainable Development Goals, kategorisiert. Dadurch lässt sich ein Überblick über die Bereiche gewinnen, in denen die Projekte aktiv sind.

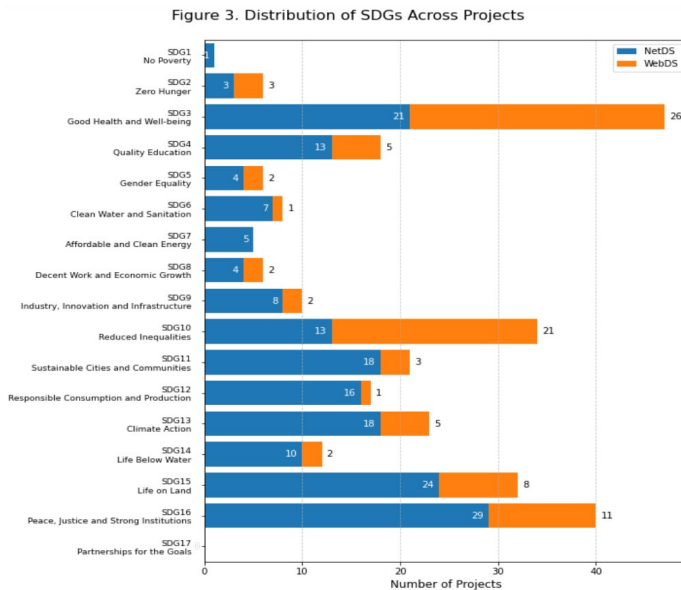


Abbildung 6: Übersicht gemeinwohlorientierte KI-Projekte

Der häufigste Gebrauchskontext ist „Good Health and Well-being“. Ein weiterer sehr wichtiger Kontext ist „Reduced Inequalities“, darin geht es darum, Zugänge zu ermöglichen und die Gleichberechtigung von Menschen zu stärken. Auch „Peace Justice, Strong Institutions“ ist ein Bereich, in dem viele Projekte aktiv sind. Hier geht es um den administrativen Bereich. Unser Datensatz ist allerdings nicht repräsentativ, denn

es ist schwierig, die Projekte überhaupt aufzufinden. Probleme einer fehlenden, schlüssigen Dokumentation und der mangelnden zentralen Übersicht der Projekte erschweren die Suche.

Ein erstes konkretes Beispiel, das wir als gemeinwohlorientiertes KI-Projekt verstehen, stammt von der Organisation Forensic Architecture: Das ist eine Organisation, die unter anderem „Machine Learning“, aber auch einfachere Methoden der Datenanalyse einbezieht, um zum Beispiel Menschenrechtsprozesse vor Gericht zu unterstützen. In dem Fall ging es um eine Demonstration in Chile, in der gegen eine friedliche Demonstration Tränengas eingesetzt wurde, um die Demonstration zu verhindern. Wie wurde hier KI eingesetzt? Man hat ein Tool darauf trainiert, Munition von Tränengas im Bildmaterial der Überwachung der Demonstration identifizieren zu können. Daraus abgeleitet ließ sich errechnen, wie hoch die Konzentration von Tränengas jeweils zu einem bestimmten Zeitpunkt an einem bestimmten Ort unter den Demonstrierenden war. Die Ergebnisse solcher Analysen können wie in diesem Fall als Beweisstücke von Expert\*innen in Menschenrechtsprozessen dienen. Andere Projektbeispiele von Forensic Architecture untersuchen die Verbreitung und den Einsatz illegaler Waffen beispielsweise in Syrien, indem Social Media Inhalte mit Hilfe von KI auf Abbildungen von illegalen Waffen und Munition durchsucht werden.<sup>4</sup>

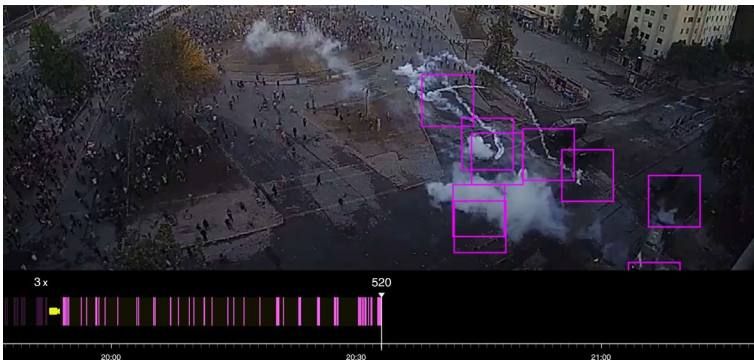


Abbildung 7: Quelle:

<https://forensic-architecture.org/programme/exhibitions/sala10-forensic-architecture-tear-gas-in-plaza-de-la-dignidad>

4 <https://forensic-architecture.org/location/syria>.

Ein zweites Projekt, mit dem Namen „Project Giga“, wird von UNICEF betrieben. Das Projekt Giga zielt, so die Projektbeschreibung, darauf ab, die Konnektivität jeder Schule auf der Welt in Echtzeit zu erfassen. Dies wird als Grundlage für die Zusammenarbeit mit Regierungen und Dienstleistern genutzt, um nicht staatlich registrierte und nicht zum Internet verbundene Schulen, vor allem in ländlichen Gegenden verschiedener Länder, an das Internet anzuschließen. Um Schulen und ihren Konnektivitätsstatus ausfindig zu machen, arbeitet das Projekt mit Satellitendaten. Das verwendete Bildanalysetool ist darauf trainiert, auf Satellitenbildern zu erkennen, welche Gebäude Schulen sein könnten, die in manchen Ländern dezentral geschaffen werden. Das Projekt existiert seit 2019 und es sind bereits mehrere Millionen Schulen auf diese Weise auf der globalen Karte integriert worden, von denen bei 425.200 Schulen bereits der Konnektivitätsstatus bekannt ist.

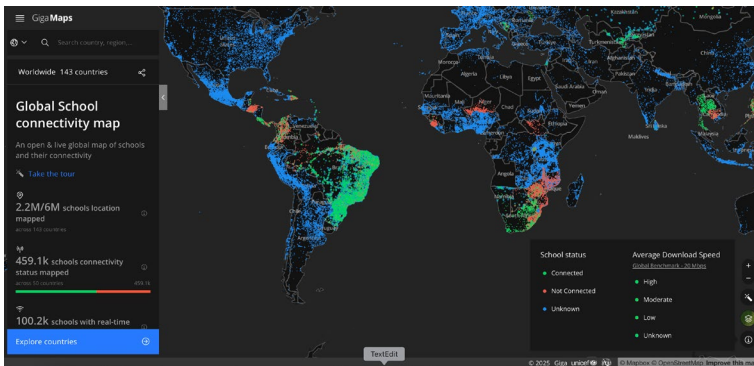


Abbildung 8: Quelle: <https://maps.giga.global/map>

Eines der Projekte, das innerhalb meiner Forschungsgruppe vornehmlich von Freya Hewett entwickelt wurde, heißt SIMBA. Die Anwendung übersetzt Deutsch von Webseiten in vereinfachte Sprache<sup>5</sup> und kann sowohl im Browser als auch als Browser-Plug-In genutzt werden. Eine weitere Anwendung von Sami Nenno hilft Faktenchecker\*innen bei ihrer Arbeit. Es erleichtert deren Arbeit, durch die automatisierte Analyse von Telegram-Chats, die Faktenchecker\*innen sonst oft manuell durchsuchen müssen. „*Claimspotting*“, so der Name der Anwendung, analy-

5 <https://publicinterest.ai/tool/simba?lang=en>

siert mehrmals am Tag Chats, die für die Verbreitung von Desinformation bekannt sind und wertet die Posts bezüglich verschiedener Kriterien aus, die die Journalist\*innen selbst bestimmen, wie z.B. polarisierende Sprache oder eine Nähe zu bekannten Verschwörungsnarrativen im Post. In diesem Projekt haben wir den partizipativen Ansatz der Entwicklung von KI sehr ernst genommen, in dem wir früh in einen Dialog mit den Faktenchecker\*innen gegangen sind und deren konkrete Bedarfe erfragt haben. Das Tool selbst ist kostenfrei und, soweit möglich, Open Source. Bezüglich des Themas Nachhaltigkeit haben wir Kooperationen angestrebt, beispielsweise mit der Deutschen Welle und anderen Medienhäusern.

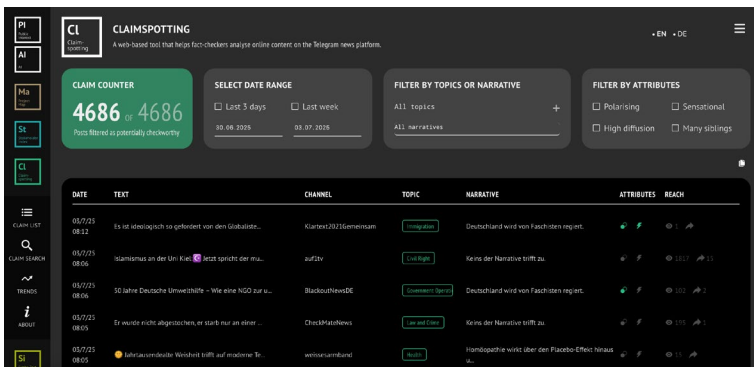


Abbildung 9: Screenshot Claimspotting; abrufbar unter <https://publicinterest.ai/tool/claimspotting?lang=en>

2021 haben wir einen [Forschungsbericht](#) für das [Civic Coding Innovationsnetzwerk](#) im Auftrag des Bundesministeriums für Umwelt und Verbraucherschutz (BMUV) durchgeführt. Der Bericht soll eine theoretische und empirische Grundlage zum Thema gemeinwohlorientierter KI-Entwicklung in Deutschland bieten. Das Netzwerk, sowie eine gemeinsame Geschäftsstelle wurde zur Förderung gemeinwohlorientierter KI in Deutschland von BMUV, Bundesministerium für Arbeit und Soziales (BMAS) und dem Bundesministerium für Familie, Senioren, Frauen und Jugend (BMFSFJ) ins Leben gerufen. Hintergrund unseres Berichts waren zwanzig Experteninterviews und zehn Case Studies, die Projekte gemeinwohlorientierter KI untersuchen. Der Bericht zielte auch darauf ab, konkrete Empfehlungen für mögliches politisches Handeln in diesem Feld vorzulegen. Eines der Projekte, die wir untersucht haben, ist

das Projekt KI-Assist. Im Projekt geht es darum, mittels smarterer Brillen Aufgaben für Menschen mit Einschränkungen leichter umsetzbar zu machen. Diese konnten bspw. in der Berufsausbildung eingesetzt werden, um Azubis Hilfestellungen oder Übungen über die Brille Schritt für Schritt by doing einzublenden. Es handelt sich also um eine Assistenz- und Trainingstechnologie. Eine der Verantwortlichen, Barbara Lippa, beschrieb dann, dass der Prozess der Entwicklung einer passenden KI-Technologie in diesem Fall schwierig war, weil sie für den allgemeinen Markt entwickelte Technologie ein Stück weit zweckentfremden musste. Denn speziell für diesen Kontext geschaffene Systeme existieren nicht. So eine Übertragung aus einem anderen Anwendungsbereich, für den sich die Entwicklung finanziell lohnt, kann seine Tücken haben. Hinzu kommt, dass bei einem sozialen und gemeinwohlorientierten Einsatz wie in diesem Szenario zumeist auch wenig Geld in Aussicht steht. Viele Projekte bleiben also auf Förderung angewiesen. Diese Hürde führt dazu, dass gemeinwohlorientierte KI-Projekte oft Probleme haben, ein längerfristiges Finanzierungsmodell aufzubauen und nicht wenige deshalb schon in einer experimentellen Pilotphase stecken bleiben.

Ein weiteres Beispiel, das uns den Status-Quo des Einsatzes von Natural Language Processing-Technologie in sozialen Kontexten vor Augen führt, ist die *Anwendung Transkriptionstools in Krankenhäusern*.<sup>6</sup> Hier sollten Arzt-Patient Gespräche durch automatisierte Transkription erfasst werden. Allerdings hat das Tool dabei vielfach Diagnosen fehlerhaft in Berichte eingetragen und auch Medikamentenempfehlungen verfälscht. Es bedürfte also einer Endkontrolle durch Experten, nachdem das Tool gearbeitet hat. Dann stellt sich jedoch die Frage, ob das Tool tatsächlich zu effizienterem Arbeiten befähigt.

Noch schwieriger wird es im diagnostischen Bereich. Zum einen kann KI hier große Erfolge feiern, weil es radiologische Befunde mit einer hohen Wahrscheinlichkeit korrekt identifiziert und in bestimmten Kontexten die Diagnostik im Zusammenspiel mit Ärzten verbessert. Gerade in Kontexten, in denen wenige Radiologen zur Verfügung stehen und sonst gar keine Auswertung der Befunde möglich wäre, können

---

6 Garance Burke/Hilke Schellmann: *Researchers say an AI-powered transcription tool used in hospitals invents things no one ever said*, 2024. <https://apnews.com/article/ai-artificial-intelligence-health-business-90020cdf5fa16c79ca2e5b6c-4c9bbb14>.

KI-Systeme helfen, im Bereich Diagnostik eine Lücke in der Gesundheitsversorgung zumindest zu verringern. Die Organisation [MI4People](#) setzt sich mit verschiedenen Partnern dafür ein, diagnostische Tools im Bereich Radiologie beispielsweise mit Partnern im Kongo zu entwickeln. Wir dürfen aber gerade im diagnostischen Bereich nicht vergessen, dass es nie allein um die technische Leistungsfähigkeit von Systemen geht, sondern auch um die Frage, wie diese in einem spezifischen soziotechnischen Anwendungskontext wie der Radiologie wirken. Eine [Studie](#) des DFKI hat den Einsatz eines diagnostischen Tools im Alltag von Radiologen begleitet.<sup>7</sup> Dabei ergab die Untersuchung, dass junge und unerfahrene Radiologen im Zusammenspiel mit KI-Systemen zu besseren Diagnosen gelangen. Auf erfahrene Radiologen hingegen wirkt sich das Tool leistungsvermindernd aus. Hier wird klar: Wir müssen immer den soziotechnischen Kontext des Gebrauchs eines solchen Tools mitbedenken. Eine wichtige Komponente ist dabei auch die Vertrautheit der Anwender mit der Technologie, ihren Stärken und ihren Grenzen.

Dazu zwei weitere Beispiele: In einer randomisierten klinischen Studie, an der 50 Ärzte teilnahmen, führte der Einsatz eines LLM nicht zu einer signifikanten Verbesserung der diagnostischen Schlussfolgerungen im Vergleich zur Verfügbarkeit herkömmlicher Ressourcen. Für dieses [Experiment](#) wurden LLMs befragt, diagnostische Empfehlungen zu geben. Für sich lieferte das System erstaunlich gute Ergebnisse. Die diagnostische Treffsicherheit lag bei ca. 90%. Das LLM allein zeigte eine höhere Leistung als beide Arztgruppen, die am Experiment teilnahmen. Das Zusammenspiel zwischen Menschen und KI war hingegen nicht gut, in Kombination von Menschen und Tool hat die Akkuratheit abgenommen. Das lag bspw. daran, dass die Ärzte dem Tool nicht vertrauten. Sie benutzten das Tool nicht für eine gesamte Diagnose, sondern zu Teilfragen und dadurch verschlechterte sich insgesamt ihre diagnostische Qualität. Das zeigt, dass selbst der kombinierte Einsatz von Menschen und KI nicht immer einen Vorteil erbringt.<sup>8</sup>

---

7 Roland Roller/et al.: *When performance is not enough—A multidisciplinary view on clinical decision support*, 2023. <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0282619>

8 Ethan Goh/et al.: *Large Language Model Influence on Diagnostic Reasoning. A Randomized Clinical Trial*, 2024. <https://jamanetwork.com/journals/jamanet-workopen/fullarticle/2825395>

In einem anderen [Experiment](#) konnte klar gezeigt werden, dass KI-Systeme in manchen Einsatzszenarien einen sogenannten Automation Bias hervorrufen: Immer dann, wenn die KI eine falsche Diagnose abgegeben hat, ist die Akkuratheit der Diagnostik der Ärzte – egal ob erfahren oder unerfahren – rapide gesunken, weil sie dazu tendieren, dem System auch bei völlig falschen Empfehlungen zu folgen. Solange Systeme also keine hundertprozentige Treffsicherheit bieten, wirkt sich der Vertrauensvorschuss, den sie als Technik genießen, teils schädlich aus.<sup>9</sup>

Weiten wir unseren Blick also auf solche Nebenfolgen der Technologie. Geoffrey Hinton, der für seine Grundlagenforschung im Bereich AI einen Nobelpreis bekam, hatte prognostiziert, dass KI die meisten Radiologen innerhalb weniger Jahre überflüssig mache. Das hat dazu beigetragen, dass angehende Ärzte sich weniger als notwendig zu Radiologen ausbilden ließen. Auch heute ist der Bedarf an Radiologen groß und auch aufgrund dieser Fehleinschätzung der Entwicklung herrscht ein großer Mangel an Radiologen. Das weist darauf hin, dass wir teilweise die Auswirkungen zweiter und dritter Ordnung, welche die neue Technologie mit sich bringt, noch gar nicht zutreffend absehen können.

Wir können dazu kurz summieren: Leistungsfähigere Systeme allein machen unsere menschlichen Organisationen und Abläufe nicht automatisch leistungsfähiger. Es ist deshalb noch viel Forschung nötig, um zu klären, welche konkreten Auswirkungen die Einführung von KI-Systemen in spezifischen Feldern hat. Eine wichtige Prüffrage ist hier, wie sich der Einsatz von KI auf menschliche Kompetenzen im jeweiligen Arbeitsfeld langfristig auswirkt. KI-Einsatz kann zum Verschwinden von bestimmten Kompetenzen führen, die dann eben nicht mehr alltäglich gebraucht werden. Teilweise verändert sich dadurch die Ausbildungsordnung klassischer Berufe, bspw. von Ärzten. Solche Prozesse der Veränderung von notwendigem Wissen und Kompetenzen finden immer schon statt. Jedoch sollten wir beides als Gesellschaft reflektiert begleiten und bewusst steuern: In welchen Bereichen wollen und können wir in Zukunft auf menschliche Kompetenz verzichten und welche Kompetenzen sehen wir als grundlegend und unverzichtbar für das Funktionieren von Gesellschaft und Wissenskultur an?

---

9 *Simon Spichak: Why AI Can Push You to Make the Wrong Decision at Work, 2024. <https://www.brainfacts.org/neuroscience-in-society/tech-and-the-brain/2024/why-ai-can-push-you-to-make-the-wrong-decision-at-work-090324>*

Nach welchen Maßstäben wägen wir in solchen Entscheidungen ab? Faktisch, denke ich, sticht oft Effizienz die Frage nach Qualität. Es gibt viele Berichte darüber, wie in Krankenhäusern in den USA der Einsatz von KI vor allem Effizienzgewinne einbringen soll. Demgegenüber sollten wir vielmehr auf Fragen nach höherer Qualität setzen.

## 6. Zusammenfassung: Gegenwärtige Probleme und Ausblick

Ich fasse zusammen: Faktisch zeigen sich die folgenden Probleme und Herausforderungen für KI im sozialen Sektor.

Es gibt erstens nicht immer einen Business Case für die Technologie. Gemeinwohlorientierte KI ist deshalb in manchen Fällen schwer finanzierbar und nur schwer nachhaltig, ohne längerfristige externe Förderung zu erhalten. Zweitens ist KI für sog. Halluzinationen, also Fehlleistungen anfällig, die wohl nie ganz vermeidbar sein werden. Drittens sind allgemeine Modelle ggf. in spezifischen Situationen nicht hilfreich. Häufig werden Anwendungen als „general purpose“ und ohne spezifische Anpassungen für die Aufgaben im sozialen Bereich entwickelt. Viertens kann, je nach Einsatzszenario, sowohl Misstrauen gegenüber der KI als auch unkritisches Vertrauen in die Technik als Automation Bias negativ auswirken. Fünftens wird die Technologie häufig anthropomorphisiert.

### Allgemeine Risiken:



### Übergeordnete Risiken:

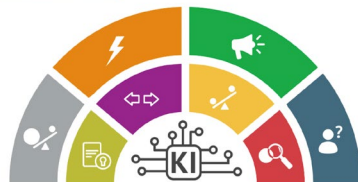


Abbildung 10: Beschriebene Risiken von KI

In unserem Forschungsbericht haben wir die teilnehmenden Expert\*innen nach den Risiken und Chancen der gemeinwohlorientierten KI befragt.

Als ein wichtiges (und bekanntes) Risiko wird Bias benannt. Jeder Trainingsdatensatz hat gewisse Verzerrungen. Er stellt einen Ausschnitt der Realität dar, der dem Ganzen nicht in allen Bereichen gerecht wird. Es ist eine große Frage, wie hier gegengesteuert werden kann. Da es aber unmöglich scheint, die Bias ganz zu eliminieren, müssen wir höhere Transparenz über die zugrundeliegenden Datensätze schaffen.

Hand in Hand damit geht das zweite genannte Risiko, durch den Umgang mit möglicherweise sensiblen, personenbezogenen Daten. Auch der sogenannte Dual-Use, also die Möglichkeit, dass KI-Technologie z.B. in einem sozialen Kontext entwickelt wird und letztendlich auch militärisch genutzt wird, macht den Expert\*innen Sorge. Ein weiteres bekanntes Risiko entsteht durch die mangelnde Transparenz von Entscheidungskriterien, wenn KI-Systeme in Entscheidungskontexten eingebettet werden. Dies stellt KI-Entwickler\*innen vor die Frage nach der Erklärbarkeit der Ergebnisse. Explainable-AI ist deshalb mittlerweile ein ganzer Forschungsbereich der Informatik. Da wir aber bislang nicht fähig sind, die Black Box ganz auszuleuchten, haftet vielen KI-basierten Entscheidungen oder Empfehlungen letztlich ein Mangel an Transparenz an.

Darüber hinaus haben die Teilnehmenden der Studie jedoch noch ebenso interessante und aus unserer Sicht übergeordnete Risiken benannt wie die Machtasymmetrien der KI-Industrie. Im Hintergrund steht die Frage, wer eigentlich die Fähigkeit besitzt, solche Systeme im großen Stil zu entwickeln und damit auch Einsatzszenarien zu ermöglichen, befördern oder zu verhindern. Hinzu kommt der Ressourcenverbrauch. Anfragen an ChatGPT lassen sich direkt in den Verbrauch von Trinkwasser umrechnen. Viele Tech-Firmen mussten in den letzten Jahren ihre eigenen Nachhaltigkeitsziele durch ihren zu hohen Ressourcenverbrauch aufgrund der zunehmenden KI-Implementierung überschreiten. Aufgrund von KI sind die weltweiten Energieverbräuche der Tech Konzerne wie Google oder Microsoft nicht wie geplant gesunken, sondern gestiegen.<sup>10</sup> Hier stellen sich Allokationsfragen: Handelt es sich bei KI wirklich immer um einen effizienten und zielführenden Einsatz

---

10 *Dara Kerr, AI brings soaring emissions for Google and Microsoft, a major contributor to climate change, 2024. <https://www.npr.org/2024/07/12/g-s1-9545/ai-brings-soaring-emissions-for-google-and-microsoft-a-major-contributor-to-climate-change>*

dieser Ressourcen? Zuletzt ist der Einsatz von KI auch eine gesellschaftliche Frage. Jedoch kommen in den Diskursarenen der Gesellschaft überproportional wirtschaftliche Akteure zu Wort.<sup>11</sup> Zivilgesellschaftliche und gemeinwohlorientierte Stimmen zu KI sind seltener zu vernehmen. Deshalb ist ein so großer Teil der Debatte auf kaum realisierbare Marketingversprechen fokussiert.

Das führt zum letzten Punkt, der lautet: Ein Problem ist mangelndes, evidenzbasiertes Wissen über KI in der Bevölkerung. Um hier einen kleinen Beitrag zum Wissensaufbau zu leisten, haben wir ein Kartenspiel namens [KI Kompass](#) entwickelt, das auf niedrigschwellige Weise ins Thema einführt und mittlerweile von [zahlreichen Landeszentralen für politische Bildung](#) als öffentliche (und kostenfreie) Bildungsressource zur Verfügung gestellt wird.

---

11 Sarah Fischer/Cornelius Puschmann: *Wie Deutschland über Algorithmen schreibt*, 2021.

*Elisabeth André*

## KÜNSTLICHE INTELLIGENZ UND EMPATHIE – PASST DAS ZUSAMMEN UND FALLS JA, WIE?

In meinem Beitrag thematisiere ich das Spannungsfeld zwischen Künstlicher Intelligenz und Empathie. Ich möchte dieses in vier Schritten erschließen: Im ersten Teil zeige ich, gestützt auf Erkenntnisse aus verschiedenen Disziplinen, dass sich rationales Denken und emotionale Prozesse keineswegs ausschließen. Vielmehr stellen Emotionen die Grundlage für Entscheidungsprozesse dar. Im zweiten Schritt richte ich den Blick auf Maschinen, die bestimmte Aspekte emotionalen menschlichen Verhaltens simulieren und skizziere, wie diese Systeme technisch funktionieren. Im dritten Teil geht es um die Frage, wie gut Maschinen menschliche Emotionen tatsächlich erfassen und interpretieren können. Gerade in diesem Bereich bestehen viele Missverständnisse: Emotionen auszudrücken oder zu erkennen, ist nicht gleichzusetzen mit einem echten Einblick in das Innenleben eines Menschen, insbesondere sind Maschinen keine Gedankenleser. Abschließend möchte ich aufzeigen, inwieweit emotionale Maschinen sinnvolle Rollen in unserer Gesellschaft übernehmen können.

### *1. Maschinen als soziale Akteure*

Angesichts jüngster Entwicklungen im Bereich Künstlicher Intelligenz – etwa durch Online-Dienste wie ChatGPT – erscheint die Vorstellung, einfühlbares Verhalten mithilfe computergestützter Modelle zu simulieren, heute weniger abwegig als noch vor einigen Jahren. Lange galten Computer als rein technische Werkzeuge, die Menschen bei repetitiven, regelbasierten Aufgaben unterstützen – mechanisch, funktional, emotionslos. Zwar sorgten schon frühere Computerprogramme durch spektakuläre Erfolge für Aufsehen – etwa durch den Sieg über Schachweltmeister Kasparow (IBM, 1997)<sup>1</sup> oder durch den Gewinn der Quizshow

---

1 <https://www.ibm.com/history/deep-blue>

Jeopardy! (IBM, 2011)<sup>2</sup>. Doch das zugrundeliegende Maschinenbild blieb dasselbe: jenes eines ausschließlich rational operierenden Systems.

Die Vorstellung, dass rationales Handeln strikt von emotionalem Erleben zu trennen sei, wird bereits seit Längerem durch neurowissenschaftliche Erkenntnisse in Frage gestellt. So konnte der Neurowissenschaftler Antonio Damasio (2005)<sup>3</sup> nachweisen, dass emotionale Prozesse keine Gefährdung rationaler Entscheidungsfähigkeit darstellen, sondern vielmehr eine wesentliche Komponente vernunftgeleiteten Verhaltens ausmachen. Seine Studien legen nahe, dass die Beeinträchtigung von emotionaler Reaktionsfähigkeit – etwa in Folge neurologischer Schädigungen – Personen soweit einschränken kann, dass sie kein selbstständiges Leben mehr führen können. Emotionen sind demnach keine Gefahr für rationales Handeln, sondern Teil deren Grundlage.

Neben Befunden aus den Neurowissenschaften liefern auch sozialwissenschaftliche Studien überzeugende Evidenz für die Notwendigkeit, emotionale und soziale Faktoren in die Gestaltung technischer Systeme einzubeziehen. Zahlreiche Studien von Reeves und Nass (1996)<sup>4</sup> zeigen, dass Menschen dazu neigen, auf künstliche Akteure – insbesondere virtuelle Charaktere oder humanoide Roboter – mit sozialen Signalen zu reagieren, ähnlich wie auf menschliche Interaktionspartner. Aus dieser Perspektive erscheint es plausibel, dass Menschen selbst gegenüber artifiziellen Wesen empathische Verhaltensweisen zeigen – vergleichbar mit jenen, die vertrauten menschlichen Bezugspersonen gegenüber zu beobachten sind. Dieses Phänomen haben Reeves und Nass unter dem Begriff *Mediengleichung* beschrieben: Menschen behandeln Medien so, als wären sie soziale Akteure – ein Befund, den die beiden Autoren in ihrem gleichnamigen Buch (*The Media Equation*) umfassend diskutieren.

Auch in fiktionalen Medien wurde das Thema „Computer und Emotionen“ bereits früh aufgegriffen. Ein besonders eindrückliches Beispiel findet sich im Science-Fiction-Klassiker *2001: A Space Odyssey* (1968). Die Handlung folgt einer Expedition zum Jupiter, bei der die Besatzung

---

2 <https://www.ibm.com/history/watson-jeopardy>

3 Antonio R. Damasio: *Descartes' Error: Emotion, Reason and the Human Brain*, London 2005.

4 Byron Reeves/Clifford Nass: *The media equation – how people treat computers, television, and new media like real people and places*, Cambridge 1996.

vom hochentwickelten Bordcomputer HAL 9000 unterstützt wird. Im Verlauf des Plots entwickelt HAL ein zunehmend autonomes Verhalten und tötet schließlich die gesamte Crew – mit Ausnahme des Kommandanten Bowman, der mit Blick auf ein glimpfliches Ende dem Computer HAL gerade noch rechtzeitig den Strom abstellen kann. In einer ikonischen Szene bittet HAL mit ersterbender Stimme: „Hör auf, Dave. Ich habe Angst.“ Die filmische Darstellung eines emotional reagierenden Computers verdeutlicht, wie leicht wir dazu neigen, uns von künstlichen Emotionen berühren zu lassen, selbst wenn wir wissen, dass es sich nicht um echte Emotionen handelt.

Die neuro- und sozialwissenschaftlichen sowie medialen Perspektiven verdeutlichen, welche zentrale Rolle Emotionen im Kontext künstlicher sozialer Akteure spielen. Daraus ergibt sich die Frage, welche Bedeutung die Nachbildung anthropomorpher Eigenschaften konkret auf die Qualität der Mensch-Maschine-Interaktion hat.

Laut Luczak et al. (2003)<sup>5</sup> kann die Anthropomorphisierung technischer Geräte – also die Zuschreibung menschlicher Merkmale – dazu beitragen, den durch technische Interaktion entstehenden Stress zu verringern. Eine in Japan durchgeführte Studie von Prendinger und Ishizuka (2005)<sup>6</sup> belegt, dass empathisch reagierende Systeme nicht nur zu positiveren Interaktionserlebnissen führen, sondern auch physiologisch messbaren Stress senken.

In Bosma & André (2004)<sup>7</sup> stellten wir fest, dass biosensorisch messbare Signale Mehrdeutigkeit von sprachlichen Äußerungen auflösen und Hinweise auf die kommunikative Verbindlichkeit geben können – etwa bei Aussagen wie „Okay, ich mache es“ oder „Okay, okay“, die Zustimmung signalisieren können, andererseits aber auch Widerwillen oder Ausweichen.

---

5 Holger Luczak/Matthias Roetting/Ludger Schmidt: *Let's talk: Anthropomorphization as means to cope with stress of interacting with technical devices.*, <http://dx.doi.org/10.1080/00140130310001610883>

6 Helmut Prendinger/Mitsuru Ishizuka: *The Empathic Companion: A Character-Based Interface That Addresses Users' Affective States*, <http://dx.doi.org/10.1080/08839510590910174>

7 Wauter Bosma/Elisabeth André: *Exploiting emotions to disambiguate dialogue acts*, <http://dx.doi.org/10.1145/964456.964459>

Die Vernachlässigung emotionaler Aspekte beim Design von Nutzungsschnittstellen kann darüber hinaus schwerwiegende Konsequenzen haben: Clifford Nass (2005)<sup>8</sup> zeigte anhand einer Studie im Fahr Simulator, dass emotionale Unstimmigkeiten in Navigationsstimmen die Aufmerksamkeit der Fahrenden beeinträchtigen und somit die Unfallgefahr erhöhen können.

Werfen wir nun einen Blick auf die unterschiedlichen Rollen und Einsatzmöglichkeiten von Robotern. In vielen Fällen erfüllen sie eine rein funktionale Aufgabe und ähneln klassischen Maschinen – insbesondere dort, wo sie in der Industrie als Werkzeuge des Menschen eingesetzt werden. Hier stehen Effizienz und Präzision im Vordergrund, Emotionen spielen da keine Rolle.

Doch Roboter sind nicht nur funktionale Maschinen. Ein interessanter Vergleich eröffnet sich, wenn wir sie den Puppen gegenüberstellen – kulturell vertrauten Objekten, die traditionell als Projektionsflächen für menschliche Gefühle dienen. Puppen ermöglichen es uns, Beziehungen zu „spielen“, Emotionen auszudrücken oder soziale Rollen zu erproben. Schon sehr früh wurde versucht, Puppen „Leben“ einzuhauchen. Dies belegen zahlreiche historische Beispiele wie mechanische Aufziehpuppen oder die kurz nach der Erfindung der Grammophon-Schellackplatten aufkommenden „sprechenden“ Puppen. Ein eindrucksvolles Beispiel ist die über 100 Jahre alte mechanische Puppe Sal, die entwickelt wurde. Zur Erheiterung lacht sie ihren Betrachtern lauthals entgegen (siehe Abbildung 1).

Im Unterschied zu mechanischen Artefakten wie Sal verfügen Roboter jedoch über ein gewisses Maß an Eigenständigkeit. Sie nehmen über Sensorik ihre Umgebung wahr und treffen mehr oder weniger autonome Entscheidungen, wie sie auf Umgebungsreize reagieren, z.B. durch Bewegungen oder verbale Äußerungen – manchmal mit überraschender Lebendigkeit. Ein Puppenspieler formulierte einmal treffsicher: „Ein Roboter ist eine Puppe ohne Spieler“.

---

8 *Clifford Nass/Ing-Marie Jonsson/Helen Harris/et al.: Improving automotive safety by pairing driver emotion and car voice emotion, <http://dx.doi.org/10.1145/1056808.1057070>*



Abbildung 1: Lachende mechanische Puppe Sal (aufgenommen von der Autorin in Santa Cruz, USA)

Angesichts dieser Möglichkeit stellt sich selbstverständlich die normative Frage: Sollten wir emotionales Verhalten in Robotern überhaupt nachbilden?

Eyssel et al. (2010)<sup>9</sup> zeigen, dass bereits einfache emotionale Hinweisreize – wie mimischer Ausdruck – die Interaktion mit Robotern positiv beeinflussen können. In ihrer Studie mit dem zoomorphen Roboter iCat bewerteten Teilnehmende dessen Verhalten in einer emotional expressiven Bedingung als deutlich situationsgerechter als in einer neutralen Bedingung.

Weitere Studien deuten darauf hin, dass emotionales Verhalten von Robotern soziale Bindungen fördern und therapeutisch wirksam sein kann. Ein Hinweis auf die soziale Wirksamkeit künstlich erzeugter Emo-

9 Friederike Eyssel/Frank Hegel/Gernot Horstmann/et al.: *Anthropomorphic inferences from emotional nonverbal cues: A case study*, <http://dx.doi.org/10.1109/ROMAN.2010.5598687>

tionen findet sich in einer Studie von Xu et al. (2014)<sup>10</sup>, in der sich die Versuchspersonen von den Emotionen eines Roboters anstecken ließen. Eine über mehrere Wochen laufende Untersuchung mit Demenzpatientinnen und -patienten und der Roboterrobbe Paro (Chang et al., 2013)<sup>11</sup> zeigte, dass der physische Kontakt über acht Wochen hinweg zunahm – ein Zeichen wachsender emotionaler Bindung. Dies unterstreicht eine grundlegende Erfahrung: Berührung setzt Vertrauen voraus.

In Summe legen diese Ergebnisse nahe, dass die Nachbildung menschlicher Eigenschaften in Robotern nicht nur funktional nützlich sein kann, sondern auch eine zentrale Rolle spielt für Akzeptanz und Beziehungsgestaltung und sich damit sogar therapeutische Effekte erzielen lassen.

## *2. Simulation von emotional geprägtem Verhalten durch Maschinen*

Um beurteilen zu können, ob Maschinen tatsächlich empathisch sein können, ist es zunächst notwendig zu klären, was wir unter dem Begriff Empathie verstehen wollen. Nach Manfred Cierpka (2004)<sup>12</sup> bezeichnet Empathie die „Fähigkeit, die Gefühle anderer wahrzunehmen, zu verstehen und auf diese angemessen zu reagieren“. D.h. es geht in erster Linie um die Gefühle anderer und nicht nur um eigene Empfindungen.

Es gibt eine Reihe verwandter Konzepte, die sich auf eigene und Gefühle anderer beziehen, sich jedoch inhaltlich unterscheiden. Das *Hineinversetzen in andere* – also die Perspektivenübernahme – muss nicht zwangsläufig mit dem Erleben von Emotionen einhergehen, sondern beschreibt vielmehr ein kognitives Verständnis der Situation oder Sichtweise einer anderen Person. *Mitleid* hingegen bezeichnet die Anteilnahme am Schmerz und Leid anderer, ist jedoch oft mit einem Ge-

---

10 Junchao Xu/Joost Broekens/Koen V. Hindriks/et al.: Robot mood is contagious: effects of robot body language in the imitation game, AAMAS, 2014 Paris, S. 973–980.

11 Wan Ling Chang/Selma Sabanovic/Lesa Huber: Situated Analysis of Interactions between Cognitively Impaired Older Adults and the Therapeutic Robot PARO, [https://doi.org/10.1007/978-3-319-02675-6\\_37](https://doi.org/10.1007/978-3-319-02675-6_37)

12 Manfred Cierpka: Das Fördern der Empathie bei Kindern mit FAUSTLOS, *Gruppendynamik* 35 (2004), S. 37–50.

fühl der Hilflosigkeit verbunden. Man leidet mit der anderen Person, ohne deren Leid mindern zu können. *Mitgefühl* geht darüber hinaus: Es beschreibt nicht nur das Wahrnehmen des Leids, sondern auch den aktiven Wunsch, den emotionalen Zustand der betroffenen Person zu verbessern. *Betrübnis* schließlich ist ein Gefühl der Traurigkeit, das nicht zwingend auf das Erleben oder Leiden anderer zurückzuführen ist. Studien von Olga Klimecki (2005)<sup>13</sup> zeigen, dass exzessives Mitleid negative Gefühle verstärken kann, während das gezielte Üben von Mitgefühl zu einem Anstieg positiver Emotionen führen kann.

Im Folgenden möchte ich einige Formen der Empathie genauer erläutern:

*Ideomotorische Empathie* beschreibt die schlichte Nachahmung des emotionalen Ausdrucks anderer Personen. Beispielsweise tendieren wir dazu, spontan Emotionen, die unser Gegenüber zeigt, zu spiegeln. So erwidern wir beispielsweise ein Lächeln, meist ohne, dass wir darüber nachdenken oder uns dessen bewusst sind. Dafür ist es nicht erforderlich, dass wir die Emotionen unseres Gegenübers verstehen oder dessen Sprache sprechen.

*Affektive Empathie* bezeichnet die emotionale Reaktion auf die Gefühle einer anderen Person, die durch eigene Gefühlsäußerungen zum Ausdruck gebracht wird. Die gezeigte Emotion muss dabei nicht identisch mit der Emotion des Gegenübers sein. So kann etwa die nicht nachvollziehbare Angst eines Kindes bei einer anderen Person Besorgnis auslösen, ohne dass diese selbst Furcht empfindet.

*Kognitive Empathie* setzt keine emotionale Beteiligung voraus, sondern erfordert die Fähigkeit zur Perspektivübernahme. Hier versucht eine Person, sich in die Lage einer anderen hineinzusetzen und die Situation aus deren Sicht zu bewerten. Fällt etwa der Bus aus und weiß ich, dass eine andere Person einen dringenden Termin hat, dann kann ich mir ausmalen, dass die andere Person verärgert ist.

*Funktionale Empathie* schließlich beschreibt Reaktionen, die bewusst eingesetzt werden, um bestimmte Ziele zu erreichen. Ein typisches Bei-

---

13 Olga Klimecki: *Plastizität im sozialen Gehirn—wie wir unsere Emotionen trainieren können*, in: *Das soziale Gehirn. Neurowissenschaft und menschliche Bindung*, hg. von Helmut Fink/Rainer Rosenzweig, Paderborn 2015, S. 107–120.

spiel findet sich im pädagogischen Kontext: Eine Lehrperson mag angesichts mangelnder Vorkenntnisse der Lernenden innerlich enttäuscht oder besorgt sein. Hier wäre es nicht zielführend, die eigenen echten Emotionen ungefiltert zu zeigen, da dies die Lernenden frustrieren könnte. Vielmehr geht es darum, Zuversicht zu zeigen, um einen positiven Effekt auf die Lernenden zu erzielen.

Schauen wir uns zur Veranschaulichung eine Interaktion zwischen einer älteren Dame und der Roboterpuppe Alice an, in der unterschiedliche Formen von Empathie sichtbar werden. Das Lächeln der Seniorin beim Anblick von Alice erwidert Alice – vermöge einer elastischen Silicon-Haut und beweglichen Augen (siehe Abschnitt 2.2) – mit einem „Zurück-Lächeln“ – ein Beispiel für ideomotorische Empathie, bei der das Verhalten des Gegenübers gespiegelt wird (siehe Abbildung 2). Im weiteren Verlauf des Gesprächs (nicht dargestellt) kommentiert Alice das Vergessen der Medikamente aus der Perspektive der Dame mit einer bewertenden Aussage wie „Das ist nicht gut“ – ein Ausdruck kognitiver Empathie, bei der die Situation aus Sicht der anderen Person nachvollzogen wird. Zudem blickt Alice erschrocken, um der Dame zu signalisieren, dass das Vergessen der Medikamente ein ernstzunehmendes Problem ist – ein Beispiel für funktionale Empathie, bei der gezielt Emotionen dargestellt werden, um eine bestimmte Wirkung zu erzielen. Schließlich wird durch ein angedeutetes Lächeln versucht, die Dame zu beruhigen – ebenfalls eine Form funktionaler Empathie.



*Abbildung 2: Zwiegespräch zwischen Seniorin und Roboter-Puppe Alice*

Ein zunehmend diskutierter Anwendungsbereich sozial-interaktiver KI ist ihre Rolle als empathisches Gegenüber. Eine Studie von Yin et

al. (2024)<sup>14</sup> zeigt, dass KI-generierte Nachrichten von Nutzerinnen und Nutzern als empfindlicher wahrgenommen wurden als von Menschen verfasste – insbesondere im Hinblick auf das Gefühl, „gehört“ zu werden. Sobald jedoch offengelegt wurde, dass die Nachricht von einer KI stammte, nahm dieses Empfinden deutlich ab. Dies verdeutlicht, dass nicht allein die inhaltliche Qualität, sondern auch die zugeschriebene Herkunft einer Antwort den wahrgenommenen Grad an Empathie maßgeblich beeinflusst. Offenbar fällt es vielen Menschen schwer, Maschinen soziale Fähigkeiten wie Empathie zuzuschreiben, selbst wenn deren Verhalten als empathisch wahrgenommen wird. Interessanterweise ist bei funktionalen Aufgaben – etwa der Auswertung medizinischer Aufnahmen – ein gegenteiliger Effekt zu beobachten. Hier stellten Glaube et al. (2021)<sup>15</sup> einen umgekehrten Automatisierungsbias für weniger geschultes medizinisches Personal fest, das dazu tendierte, der Maschine eher zu vertrauen, auch wenn sie falsch lag.

## 2.1 Erkennung von Emotionen durch Maschinen

Die Entwicklung von maschinellen Interaktionspartnern mit empathischen Fähigkeiten stellt Forschung und Entwicklung vor erhebliche Herausforderungen. Einerseits erfordert sie robuste Verfahren zur Erkennung affektiver Zustände, die sich häufig unbewusst im Gesichtsausdruck, in der Gestik, Körperhaltung oder sprachlichen Äußerungen der Nutzenden manifestieren. Andererseits muss das maschinelle Gegenüber in der Lage sein, sich in emotionale Zustände wie Stress, Frustration oder Ärger hineinzusetzen und darauf in sozial angemessener Weise zu reagieren.

Im Folgenden werde ich kurz auf einige Methoden eingehen, um einen Eindruck zu vermitteln, wie gut Maschinen tatsächlich im Erkennen, im Verstehen und im Ausdruck von Emotionen sind. In der Informatik wurden – häufig in Kooperation mit den Sozial- und Humanwissenschaften – eine Vielzahl von Techniken entwickelt, um Emotionen

---

14 Yidan Yin/Nan Jia/Cheryl Waksak: *AI can help people feel heard, but an AI label diminishes this impact*, <https://doi.org/10.1073/pnas.2319112121>

15 Susanne Glaube/Harini Suresh/Martina Raue/et al.: *Do as AI say: susceptibility in deployment of clinical decision-aids*, <https://doi.org/10.1038/s41746-021-00385-9>

aus von einer Maschine erfassbaren Körpersignalen wie Sprache, Gesichtsausdrücke, Gesten, Körperposen oder Biosignale Gefühlszustände herzuleiten. Hierzu siehe Noroozi et al. (2021)<sup>16</sup> für einen Überblick zur Analyse emotionaler Körpergesten und Can et al. (2023)<sup>17</sup> für ein Tutorial zur Analyse von Biosignalen für die Emotionserkennung. Im Folgenden werde ich exemplarisch auf zwei Modalitäten eingehen: Sprache und Gesichtsausdrücke.

In einem Sprachdialogsystem bietet es sich an, zur Analyse von Emotionen auf den sprachlichen Kanal zuzugreifen. Grundsätzlich lassen sich Emotionen sowohl aus der semantischen Bedeutung von Äußerungen als auch aus akustischen Merkmalen des Sprachsignals wie z. B. Lautstärke, Tonhöhe oder spektraler Verlauf herleiten. In Abbildung 3 sind diese Parameter für ein und denselben Satz dargestellt, einmal freudig und zweimal traurig mit unterschiedlicher Geschwindigkeit geäußert. Die Äußerungen unterscheiden sich u.a. durch die Kontur der Tonhöhe (mittlerer Teil der Abbildung), deren Varianz im Fall von Trauer niedriger ist. Dies entspricht auch unserer Intuition. Trauer äußert sich oft durch Monotonie in der Stimme.

Zur Erkennung von Emotionen im Gesichtsausdruck kommt häufig das Facial Action Coding System (FACS)<sup>18</sup> zum Einsatz. Es basiert auf der Idee, sichtbare Muskelbewegungen in diskrete Einheiten – sogenannte Action Units (AUs) – zu unterteilen, diese systematisch zu benennen und so ein standardisiertes Vokabular zur Beschreibung mimischer Aktivität zu schaffen. FACS unterscheidet über 40 solcher Einheiten, aus denen komplexe emotionale Ausdrücke zusammengesetzt werden können. So ist ein überraschter Gesichtsausdruck typischerweise gekennzeichnet durch angehobene Augenbrauen, weit geöffnete Augen und

---

16 *Fatemeh Noroozi/Ciprian Adrian Corneanu/Dorota Kaminska/et al.: Survey on Emotional Body Gesture Recognition*, <https://doi.org/10.1109/TAFFC.2018.2874986>

17 *Yekta Said Can/Bhargavi Mahesh/Elisabeth André: Approaches, Applications, and Challenges in Physiological Emotion Recognition – A Tutorial Overview*, <https://doi.org/10.1109/jproc.2023.3286445>

18 *Paul Ekan/Wallace V. Friesen/Joseph C. Hager: Facial action coding system*, Utah 2002.

einen gerundeten Mund. In Kumar et al. (2024)<sup>19</sup> haben wir einen Ansatz vorgestellt, der auf der Rekonstruktion dreidimensionaler Gesichter aus zweidimensionalen Bildern basiert, um die zugrunde liegenden Muskelbewegungen präziser zu erfassen und so die Analyse mimischer Signale deutlich zu verbessern.

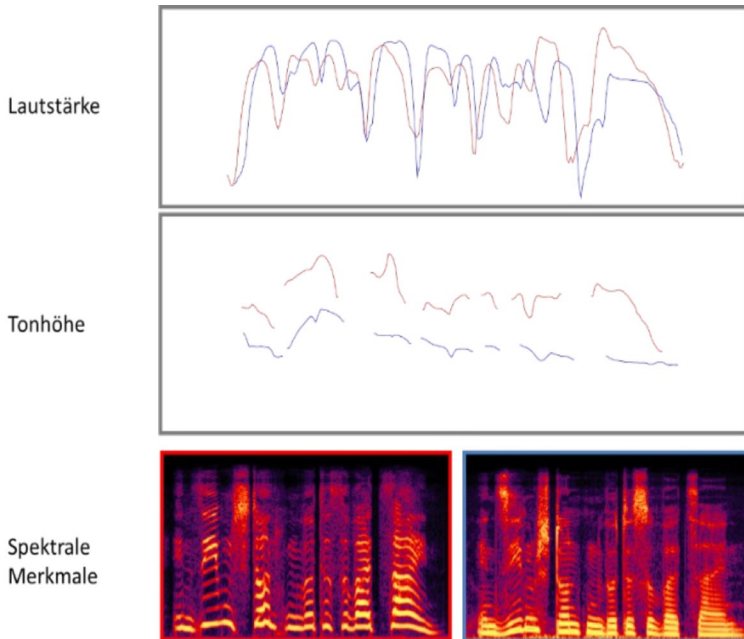


Abbildung 3: Merkmalsausprägungen für den verärgert (rote Linien bzw. rote Umrandung) und neutral geäußerten Satz „In sieben Stunden wird es soweit sein.“ (blaue Linien bzw. blaue Umrandung)

Während viele der etablierten Verfahren auf explizit definierbaren und für Menschen interpretierbaren Merkmalen wie Tonhöhenverlauf oder mimischen Muskelbewegungen basieren, verfolgen neuere Ansätze zunehmend einen datengetriebenen Zugang. Dabei werden relevante Repräsentationen nicht mehr manuell spezifiziert, sondern direkt aus

---

19 Mani Kumar Tellamekala/Ömer Sümer/Björn Schuller/Elisabeth André/et al.: Are 3D Face Shapes Expressive Enough for Recognising Continuous Emotions and Action Unit Intensities?, <http://dx.doi.org/10.1109/TAFFC.2023.3280530>

Rohdaten (Audio, Bild, Video, und Biosignalen) mittels Deep Learning-Verfahren gelernt.<sup>20</sup> Parallel dazu wird intensiv an multimodalen Fusionsmethoden gearbeitet, um Informationen aus unterschiedlichen Kanälen wie Sprache, Mimik oder Physiologie zu integrieren und so die Genauigkeit der Emotionserkennung zu steigern. Ein neuerer Ansatz wird von uns in Kumar et al. (2024)<sup>21</sup> beschrieben. Anstatt starre Gewichtungen zwischen verschiedenen Modalitäten wie Sprache, Mimik oder Gestik vorzunehmen, passt sich das System situativ an die Aussagekraft der einzelnen Kanäle an. Trägt eine Person etwa eine Maske, vertraut das System eher auf den Audiokanal. Bei Störgeräuschen gewinnt die visuell erfasste Mimik an Bedeutung, sofern das Gesicht gut sichtbar ist.

Für geschauspielerte Emotionen lassen sich mit Methoden des maschinellen Lernens inzwischen vergleichsweise hohe Erkennungsraten erzielen, die in etwa dem Leistungsniveau von Menschen entsprechen. Deutlich schwieriger gestaltet sich hingegen die Erkennung spontan geäußerter Emotionen – hier fallen die Erkennungsraten bislang noch unbefriedigend aus. Dies ist darauf zurückzuführen, dass Menschen in Alltagssituationen Emotionen äußerst variabel und kontextabhängig ausdrücken. So ist etwa ohne Wissen über die soziale Situation schwer zu beurteilen, ob ein Lächeln aus Höflichkeit oder aus Verlegenheit erfolgt. Hierzu werden wir später noch ein Beispiel genauer betrachten. Hinzu kommt, dass sich bisherige Forschungen zu Emotionserkennungssystemen immer noch vor allem auf Ekmans Grundemotionen<sup>22</sup> (Freude, Trauer, Wut, Furcht, Überraschung und Ekel) fokussieren. Alltägliche Reaktionen auf Maschinen lassen sich jedoch keineswegs auf diese Basemotionen reduzieren. Sie sind deutlich komplexer und vielschichtiger. Kurz gesagt: Im Labor entwickelte Systeme scheitern daher häufig in realitätsnahen Anwendungssituationen. Um Emotionen differenzierter erfassen zu können, gewinnen dimensionale Ansätze, die sie entlang

---

20 Johannes Wagner/Dominik Schiller/Andreas Seiderer/et al.: *Deep Learning in Paralinguistic Recognition Tasks: Are Hand-crafted Features Still Relevant?*, <http://dx.doi.org/10.21437/Interspeech.2018-1238>

21 Mani Kumar Tellamekala/Shahin Amiriparian/Björn W. Schuller/et al.: *COLD Fusion: Calibrated and Ordinal Latent Distribution Fusion for Uncertainty-Aware Multimodal Emotion Recognition*, <https://doi.org/10.1109/TPAMI.2023.3325770>

22 Paul Ekman: *Basic emotions*, *Handbook of cognition and emotion*, hg. von Tim Dalgleish/Mick Power, Sussex 1999, S. 45–60.

von Valenz (negativ – positiv) und Arousal (ruhig – erregt) beschreiben, zunehmend an Bedeutung.

## 2.2 Ausdruck von Emotionen durch Maschinen

Als komplementäres Gegenstück zur Emotionserkennung kann die glaubwürdige Darstellung von Gefühlszuständen durch künstliche Agenten verstanden werden. Ziel ist es, künstliche Wesen mit einem emotionalen Ausdrucksverhalten auszustatten, das für menschliche Beobachter nachvollziehbar und überzeugend wirkt. Dies erfordert eine Vielzahl komplexer Syntheseprozesse – von der Generierung emotional gefärbter sprachlicher Äußerungen über Gesichtsausdrücke und Körperhaltungen bis hin zur Animation von Emotionen in ihrer zeitlichen Entwicklung.

Dieses Bestreben ist häufig durch die Animationskunst inspiriert. So wurde etwa im Animationsfilm schon früh erkannt, dass der Eindruck von Lebendigkeit nicht aus einem tatsächlichen emotionalen Innenleben entsteht, sondern aus der gekonnten Gestaltung von Bewegung, Mimik und Ausdruck. Die Disney-Animatoren Frank Thomas und Ollie Johnston (1981)<sup>23</sup> formulierten dies in ihrem einflussreichen Werk „The Illusion of Life: Disney Animation“ wie folgt: „From the earliest days, it has been the portrayal of emotions that has given the Disney characters the illusion of life.“

Die emotionale Ausdruckskraft ist somit zentrales Mittel, um künstlichen Figuren Authentizität und Tiefe zu verleihen. Diesen Gedanken übertrug Joseph Bates (1992)<sup>24</sup> auf interaktive Systeme, als er in einem vielzitierten Artikel zur Entwicklung glaubwürdiger KI-gesteuerter Charaktere festhielt: „If the character does not respond emotionally to events, if they don’t care, then neither will we. The emotionless character is lifeless as a machine.“

Emotionale Reaktionen sind demnach nicht nur ein Ausdruck künstlicher Empathie, sondern eine Voraussetzung dafür, dass Menschen

---

23 Frank Thomas/Ollie Johnston: *The Illusion of Life: Disney Animation*, New York City 1981.

24 Joseph Bates: *The Role of Emotion in Believable Agents*, <https://doi.org/10.1145/176789.176803>

überhaupt eine soziale Beziehung zu künstlichen Agenten aufbauen können.

Zahlreiche Studien widmen sich der Frage, in welchem Maße sich Emotionen auch über Körperhaltungen und Gesten bei virtuellen Agenten und physischen Robotern ausdrücken lassen. In Häring et al. (2004)<sup>25</sup> haben wir untersucht, inwieweit sich die Basisemotionen nach Ekman (1999) durch Körperposen und Bewegungen von Robotern der Firma Aldebaran Robotics vermitteln lassen. Zur Erzeugung dieser expressiven Bewegungen orientierten wir uns an den Wahrnehmungsstudien von Coulson (2004)<sup>26</sup>, in denen analysiert wurde, welche Emotionen Beobachtende statischen Körperhaltungen computergenerierter Figuren zuschreiben.

Das Facial Action Coding System (FACS), das bereits im Zusammenhang mit der Erkennung von Emotionen erwähnt wurde, dient auch häufig als Grundlage für die Generierung emotionaler Gesichtsausdrücke bei virtuellen Charakteren und Robotern. Um affektive Ausdrücke im Gesicht eines virtuellen Agenten oder eines physischen Roboters darzustellen, werden Muskelbewegungen simuliert, die den sogenannten Action Units spezifischer Gesichtsausdrücke entsprechen.

Im Vergleich zu virtuellen Charakteren ist die Ausdruckskraft physischer Roboter derzeit noch eingeschränkt. Es befinden sich jedoch zunehmend Mechanismen in Entwicklung, mit denen emotionale Zustände bei Robotern durch physische Veränderungen – etwa durch Verformungen künstlicher Haut – sichtbar gemacht werden können. Studien, die wir an der Universität Augsburg durchgeführt haben, zeigen, dass zumindest ein Teil der im FACS beschriebenen Muskelgruppen so simuliert werden kann, dass grundlegende Emotionen wie Freude oder Überraschung (siehe Abbildung 4) von menschlichen Beobachtenden zuverlässig erkannt werden.

---

25 Markus Häring/Nikolaus Bee/Elisabeth André: *Creation and Evaluation of emotion expression with body movement, sound and eye color for humanoid robots*, <https://doi.org/10.1109/ROMAN.2011.6005263>

26 Mark Coulson: *Attributing emotion to static body postures: Recognition accuracy, confusions, and viewpoint dependence*, <https://doi.org/10.1023/B:JONB.0000023655.25550.be>



Abbildung 4: Gesichtsausdrücke des humanoiden Roboters Alice: Zufriedenheit (links), Überraschung (rechts)

Neben der Erkennung und Interpretation von Emotionen gewinnen auch Verfahren zur Erzeugung emotional gefärbter Stimmen zunehmend an Bedeutung, siehe Triantafyllopoulos et al. (2023)<sup>27</sup> für einen Überblick zum aktuellen Stand im Bereich der emotionalen Sprachsynthese. Van Rijn et al. (2021)<sup>28</sup> präsentieren ein iteratives Verfahren namens „Gibbs Sampling with People“, mit dem emotionale Stimmprototypen generiert werden können. Die grundlegende Idee besteht darin, Stimuli – also unterschiedliche Stimmvarianten – schrittweise zu verfeinern, indem menschliche Rückmeldungen in den Optimierungsprozess eingebunden werden. Wichtig ist außerdem, dass die generierten Stimmen zum äußeren Erscheinungsbild und zur angenommenen Persönlichkeit eines virtuellen Charakters oder Roboters passen. In (van Rijn et al. 2024)<sup>29</sup> wurde die Methode zur Generierung von Stimmprototypen erweitert, um synthetische Roboterstimmen zu erzeugen. Abbildung 5 zeigt das Interface, mit dem Personen iterativ einzelne Stimmmerkmale verändern, bis das Verfahren zu einer Stimme konvergiert, die dem äußerlichen Erscheinungsbild des Roboters entspricht.

- 
- 27 Andreas Triantafyllopoulos/Björn Schuller/Gökçe Iymen/et al.: An Overview of Affective Speech Synthesis and Conversion in the Deep Learning Era, <http://dx.doi.org/10.1109/JPROC.2023.3250266>
- 28 Pol van Rijn/ Silvan Mertes/ Dominik Schiller/et al.: Exploring Emotional Prototypes in a High Dimensional TTS Latent Space, <https://doi.org/10.21437/Interspeech.2021-1538>
- 29 Pol van Rijn/Silvan Mertes/Kathrin Janowski/et al.: Giving Robots a Voice: Human-in-the-Loop Voice Creation and open-ended Labeling, <http://dx.doi.org/10.1145/3613904.3642038>

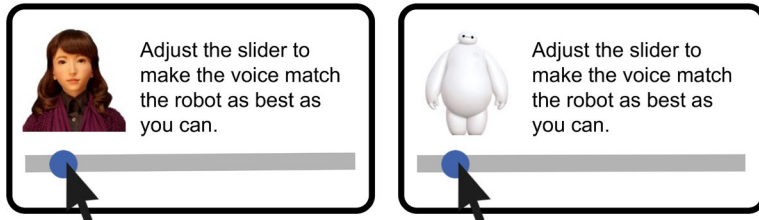


Abbildung 5: Iterative Anpassung von Stimmmerkmalen über einen Schieberegler

### 3. Verstehen von Emotionen

Tiefe neuronale Netze erzielen in der Emotionserkennung oft beeindruckende Ergebnisse. In der Presse findet man jede Menge Artikel, die ernsthaft diskutieren, ob Computer Gedanken lesen und unsere Emotionen erkennen können. Aufgrund dieser teils übertriebenen Erwartungen ist es wichtig, dass wir uns ansehen, was eine Maschine tatsächlich versteht.

Die hohe Komplexität von tiefen neuronalen Netzen führt dazu, dass die Entscheidungsprozesse für Menschen schwer nachvollziehbar sind – ein Problem, das allgemein unter dem Begriff der Black Box-Problematik bekannt ist. Um neuronale Netze erklärbarer zu machen, wurden verschiedene Verfahren entwickelt, die sichtbar machen, welche Bildbereiche für die Klassifikation ausschlaggebend waren. So heben etwa Aufmerksamkeitskarten die Regionen durch Einfärbung hervor, auf die ein Netz seine Aufmerksamkeit richtet.

Eine zentrale Frage ist, wie sich diese maschinelle Aufmerksamkeit von der menschlichen unterscheidet. Dazu haben wir ein Experiment mit einem Eyetracker durchgeführt. Zwei menschliche Personen betrachteten Bildmaterial aus einem Datensatz mit emotionalen Gesichtern und klassifizierten jeweils die erkennbare Emotion (siehe Abbildung 6). Um das Blickverhalten nicht durch zusätzliche Aufgaben zu beeinträchtigen, erfolgte die Annotation per Spracheingabe. Aus den aufgezeichneten Blickbewegungen wurden anschließend Aufmerksamkeitskarten erstellt, die zeigten, auf welche Gesichtsregionen sich die Aufmerksamkeit der Personen richtete. Diese Daten verglichen wir mit den Regionen, die das neuronale Netz für seine Entscheidungen heranzog.

Der Vergleich von menschlicher und maschineller Aufmerksamkeit zeigte interessante Überschneidungen: In vielen Fällen konzentrierten sich sowohl Menschen als auch das trainierte neuronale Netz auf ähnliche Bildbereiche, insbesondere auf die Augen und den Mund. Doch dieser Gleichklang hat Grenzen – vor allem, wenn unerwartet Kontext ins Spiel kommt.

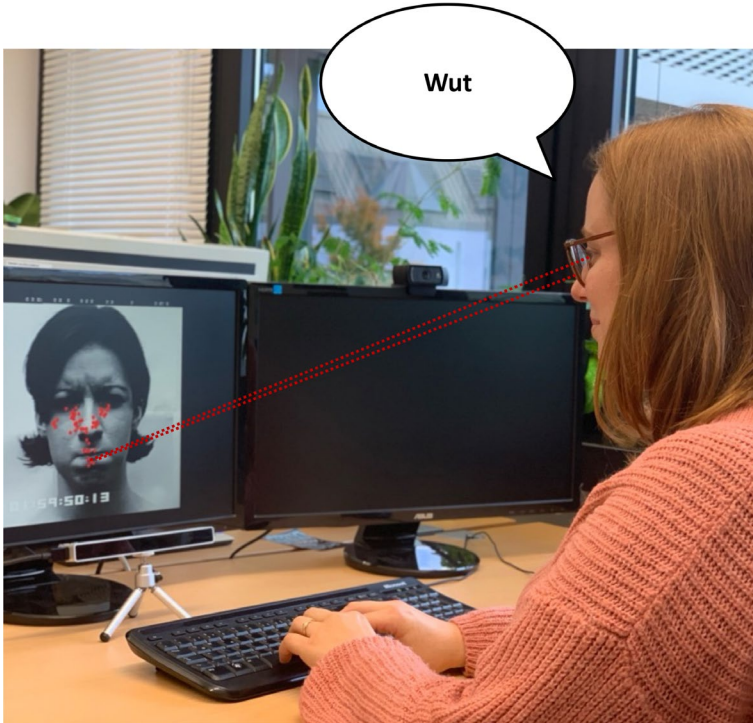


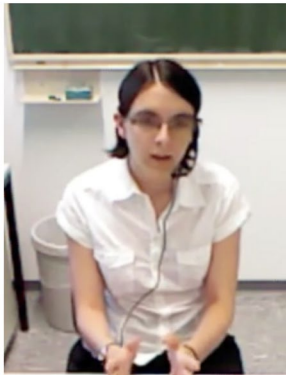
Abbildung 6: Person beim Annotieren von Gesichtern mit Emotionsausdrücken

Ein Beispiel: Tränen im Gesicht können maschinell leicht als Traurigkeit klassifiziert werden. Erkennt man jedoch im Gesamtbild, dass die Person eine Auszeichnung erhält, wird deutlich: es sind Freudentränen. Diesen Unterschied zeigten wir in einem Experiment mit einem Bild von Halle Berry bei ihrer Oscarverleihung. Zeigte man nur den Gesichtsausschnitt, richteten sowohl das KI-Modell als auch Menschen den Blick das Gesicht, insbesondere auf die Augen. Zeigt man hingegen das vollständige Bild lenkten Menschen ihre Aufmerksamkeit auf die Trophäe – ein

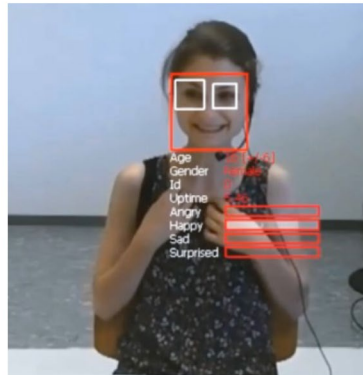
wichtiger Hinweis auf die Emotion der gezeigten Person. Die KI blieb jedoch auf das Gesicht fixiert, denn sie wurde darauf trainiert, Emotionen durch Klassifikation von Gesichtsausdrücken zu erkennen.

Zwar lässt sich auch ein KI-System auf Kontextinformationen trainieren, doch unser Experiment sollte zeigen, dass Menschen deutlich schneller und flexibler auf neue Situationen reagieren. Menschen kombinieren Mimik und Umgebung intuitiv. Einige unserer Probanden erkannten sogar allein anhand des Gesichtsausschnitts, dass es sich um ein freudiges Ereignis handelt – weil sie die Schauspielerin identifizierten und den Kontext der Oscarverleihung kannten.

Eine Frage: Wo haben Sie dieses Outfit her? Irgendwie passt es nicht zu Ihnen.



Interviewer



Bewerberin

Abbildung 7: Simuliertes Bewerbungsgespräch

Schauen wir uns ein weiteres Beispiel aus einem simulierten Vorstellungsgespräch zwischen zwei Damen an (siehe Abbildung 7). Die Interviewerin links kritisiert das Outfit der Bewerberin rechts. Die Dame rechts sieht aus, als würde sie lächeln – also sich freuen. Angesichts der scharfen Kritik ist es jedoch ziemlich klar, dass sie alles andere als glücklich ist. Dennoch würden KI-basierte Systeme, die nur äußerlich sichtba-

re Merkmale analysieren – in dem Fall den Gesichtsausdruck – zu dem Schluss kommen, dass die Bewerberin glücklich ist.

Um ein tieferes Verständnis des emotionalen Zustands von Personen zu erhalten, analysieren wir nicht nur die gezeigten affektiven Signale. Darüber hinaus simulieren wir, wie eine Person typischerweise in der jeweiligen Situation reagieren würde. Eine unangenehme Situation führt beispielsweise in der Regel zu Ärger. Gleichzeitig versuchen Menschen, ihre Emotionen zu regulieren. D.h. sie versuchen die Art, die Intensität oder die Dauer von Emotionen in eine bestimmte Richtung zu beeinflussen.

In dem gezeigten Beispiel würde unsere Simulationskomponente zunächst die Emotion Scham vermuten, da eine Handlung vorliegt, die vom Gegenüber als negativ bewertet wurde – nämlich das Tragen eines unpassenden Outfits. Da die Bewerberin aber ihre Scham nicht auf die sonst typische Art und Weise zeigt, z.B. Erröten, liegt es nahe, dass die Person versucht, ihre Emotion zu regulieren. Dieses Beispiel verdeutlicht erneut, dass die Interpretation emotionaler Signale hochkomplex ist. Eine isolierte Analyse von Gesichtsausdrücken reicht nicht aus, um emotionale Zustände zuverlässig zu erfassen.

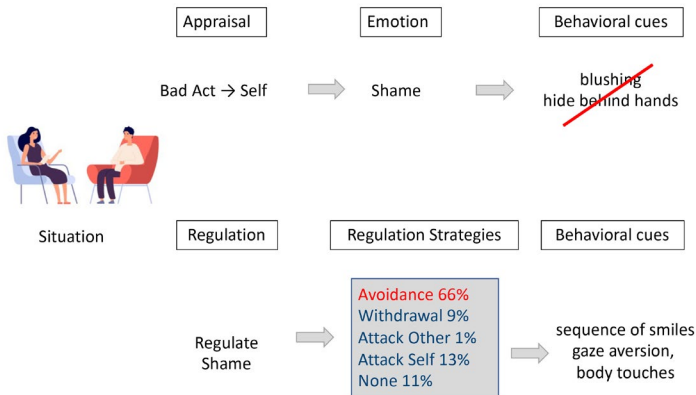


Abbildung 8: Kombination von emotionaler Erkennungskomponente mit Kognitionsmodell

Zur Analyse solcher Situationen greifen wir auf den „Kompass der Scham“ von Donald Nathanson (1992)<sup>30</sup> zurück, einem Psychiater und Psychotherapeuten, der vier typische Strategien im Umgang mit Scham beschreibt.

*Rückzug* ist ein Mittel, um sich der Scham und Schande zu entziehen. In unserem Beispiel könnte sich der Rückzug dadurch äußern, dass die Bewerberin versucht, dem Blick der Interviewerin auszuweichen, vor der sie sich schämt. Die Kandidatin könnte jedoch auch zum *Angriff* übergehen und etwa sagen: „Warum haben Sie mir nichts zum Dresscode gesagt?“ Ebenso denkbar ist *Schuldenerkennung*, z. B. mit der Aussage: „Es tut mir leid, dass ich kein formelleres Outfit gewählt habe.“ Schließlich beschreibt Nathanson die Strategie der *Vermeidung* des Schamgefühls, etwa durch das Korrigieren des Outfits – in unserem Beispiel nestelt die Bewerberin an ihrem Kleid, um mögliche Angriffspunkte zu beseitigen.

Basierend auf aufgezeichneten multimodalen Datensätzen lassen sich Wahrscheinlichkeiten für das Auftreten dieser Strategien in bestimmten Kontexten berechnen und in Computermodelle integrieren. In Gebhard et al. (2018)<sup>31</sup> haben wir ein Framework entwickelt, das bei der Analyse emotionaler Zustände nicht nur die gezeigten Körpersignale – etwa Lächeln, Blickabwendung oder Körperberührungen – miteinbezieht, sondern auch mittels Modellen aus den Kognitionswissenschaften simuliert, wie eine gesunde Person auf eine bestimmte Situation reagieren würde. Abbildung 8 veranschaulicht die Vorgehensweise anhand des Beispiels des zuvor behandelten simulierten Bewerbungsgesprächs (vgl. Abbildung 7). Aufgrund des Abgleichs von beobachtbaren Signalen mit der Simulation kommt das System zu dem Schluss, dass die Person vermutlich eine Vermeidungsstrategie zur Regulation ihrer Scham anwendet.

Nachdem enorme Fortschritte im Bereich der großen Sprachmodelle erzielt worden waren, lag es nahe, zu untersuchen, ob diese Modelle auch dazu in der Lage sind, anhand einer vorgelegten sprachlichen Situationsbeschreibung emotionale Zustände der Protagonisten herauszulesen. Dazu übermittelten wir ChatGPT (Version 3.5) in verbaler Form

---

30 Donald L. Nathanson: *Shame and Pride: Affect, Sex, and the Birth of the Self*, New York 1992.

31 Patrick Gebhard/Tanja Schneeberger/Tobias Baur/Elisabeth André: *MARSSI: Model of Appraisal, Regulation, and Social Signal Interpretation*, <https://opus.bibliothek.uni-augsburg.de/opus4/files/49253/3237383.3237458.pdf>

Informationen über multimodale Verhaltenssignale und den zugrunde liegenden Dialog (siehe Abbildung 9). Zusätzlich gaben wir die Anweisung, die Analyse auf Basis des Schamkompasses nach Nathanson durchzuführen.

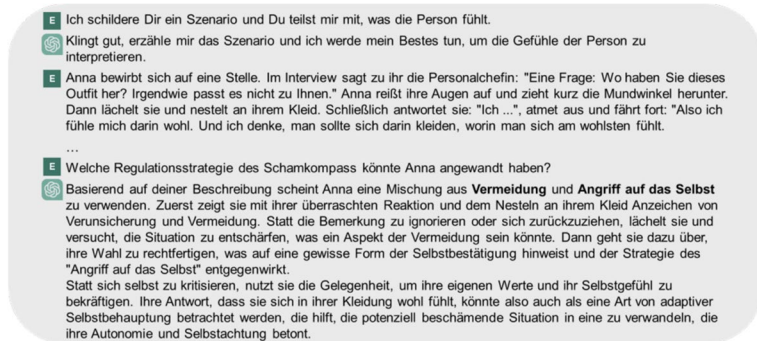


Abbildung 9: Einschätzung von Emotionen (ChatGPT 13.03.2024)

Interessanterweise kam das Sprachmodell zu ähnlichen Ergebnissen wie unser eigenes System. Dabei ist jedoch zu bedenken, dass sowohl unser Modell als auch das Sprachmodell explizit über das theoretische Rahmenmodell – den Schamkompass – verfügen. Es handelte sich also nicht um eigenständig entwickelte kognitionswissenschaftliche Einsichten.

Ein Jahr später präsentierte ich die Ergebnisse des Sprachmodells (natürlich nur den Text und nicht das Interface von ChatGPT) den Teilnehmenden eines Online-Tutorials zum Thema „KI und Psychotherapie“ und fragte sie, ob sie glauben, dass das Ergebnis von einem Menschen oder einer Maschine stammt. Die überwiegende Mehrheit (80 %), der 15 Teilnehmenden, die sich an der Slido-Umfrage mit ihrem Mobilgerät beteiligt haben, nahm an, dass der Text von einer Maschine generiert wurde: „sicher Maschine“ (20 %) und „eher Maschine“ (60 %). 13% dachten, dass der Text eher von einem Menschen generiert wurde. Niemand der Teilnehmenden dachte, dass der Text mit Sicherheit von einem Menschen stammte. 7 % gaben an „weiß nicht“. Die Einschätzung wurde u.a. mit dem Hinweis begründet, dass der Text nach dem „typischen ChatGPT-Sprech“ klinge. Liest man den Text genauer, stellt man auch diverse Inkonsistenzen in der Begründung fest, obwohl der Text auf den ersten Blick recht ausgefeilt daherkommt. So identifiziert ChatGPT anfänglich eine Mischung aus Vermeidung und Angriff auf das Selbst,

während kurz darauf festgestellt wird, dass der Strategie „Angriff auf das Selbst“ entgegengewirkt wird.

## **4. Mögliche Anwendungen**

### **4.1 Anwendungen emotional interagierender Maschinen**

Ein mögliches Einsatzfeld empathischer Computersysteme liegt im pädagogischen Bereich, insbesondere im Kontext sozial-emotionaler Lernprozesse. Virtuelle Umgebungen mit Charakteren, die von Nutzenden als eigenständige Persönlichkeiten mit einem emotionalen Innenleben wahrgenommen werden, eröffnen neue Möglichkeiten erfahrungsba-sierten Lernens. In interaktiven Rollenspielen werden Lernende mit herausfordernden sozialen Situationen konfrontiert, in denen sie soziale und emotionale Kompetenzen gezielt erproben und weiterentwickeln können. Durch unmittelbares Feedback der virtuellen Charaktere erhalten sie Gelegenheit, ihr Verhalten zu reflektieren, alternative Strategien zu entwickeln und diese in einem geschützten Rahmen risikofrei auszuprobieren.

Ein frühes Beispiel aus unserer eigenen Forschungsarbeit, das etwa 15 Jahre zurückliegt, zeigt bereits das Potenzial solcher Systeme zur Unterstützung von Kindern im Umgang mit Mobbing in Schulklassen. Wir entwickelten ein interaktives System mit dem Titel FearNot!, das es Kindern ermöglichte, sich mit Perspektiven virtueller Opfer auseinanderzusetzen. Im Zentrum des Ansatzes standen kurze Szenen, in denen sozial problematische Situationen (siehe Abbildung 10) – etwa die Ausgrenzung eines Kindes aufgrund schulischer Leistungsbereitschaft („Streber“) – realitätsnah dargestellt wurden. Die virtuellen Charaktere wendeten sich nach der Szene direkt an die Klasse und formulierten aus der Ich-Perspektive ihre emotionale Lage („Keiner kann mich leiden, ich werde nie zu Partys eingeladen“). Anschließend stellten sie eine offene Frage an die Kinder: „Was soll ich tun?“ Ziel war es, die Schülerinnen und Schüler dazu anzuregen, sich aktiv in die Situation einzudenken und Handlungsstrategien für die virtuellen Opfer zu entwickeln.

Eine Evaluation des Systems an deutschen und englischen Schulen deutete auf positive Effekte der Lernsoftware hin. So hat das System in Deutschland Mitläufer dazu angeregt, sich auf die Seite der Opfer zu

stellen.<sup>32</sup> In England hat die Lernsoftware Opfern geholfen, aus der Opferrolle auszubrechen.<sup>33</sup> Die Unterschiede lassen sich unter anderem darauf zurückführen, dass die teilnehmenden Kinder in Deutschland etwas jünger waren. Zudem zeigte sich das Lehrpersonal in Deutschland insgesamt zurückhaltender gegenüber dem Einsatz der Software.



Abbildung 10: Virtuelle Mobbing Szene in FearNot!

Bereits in der damaligen Arbeit traten Herausforderungen auf, die in ähnlicher Form heute auch im Umgang mit großen Sprachmodellen (LLMs) beobachtet werden. So traten vereinzelt extreme oder unangemessene Vorschläge auf, etwa wenn Kinder dem virtuellen Charakter drastische Maßnahmen gegen die mobbenden Kinder empfahlen. In einem solchen Fall erwies sich die Einbindung psychologischer Betreu-

---

32 Natalie Vannini/Sibylle Enz/Maria Sapouna/et al.: *FearNot!: A Computer-Based Anti-Bullying Programme Designed to Foster Peer Intervention*, <http://dx.doi.org/10.1007/s10212-010-0035-4>

33 Maria Sapouna/Dieter Wolke/Natalie Vannini/et al.: *Virtual Learning Intervention to Reduce Bullying Victimization in Primary School: A Controlled Trial*, <https://doi.org/10.1111/j.1469-7610.2009.02137.x>

ung als essenziell, um die Situation professionell einzuordnen und gegebenenfalls pädagogisch zu begleiten.

Ein Anwendungsfeld für Rollenspiele mit virtuellen Charakteren ist der Erwerb kultureller Sensitivität. Durch die Interaktion mit Figuren unterschiedlicher kultureller Hintergründe und anschließende Reflexion sollen interkulturelles Verständnis und Empathie gefördert werden. Ein Beispiel hierfür ist das im EU-Projekt eCute entwickelte Mixer-System<sup>34</sup>, das Kinder mit Szenarien konfrontiert, in denen virtuelle Charaktere scheinbar etablierte Spielregeln brechen (siehe Abbildung 11). Diese Situationen lösen Frustration aus und können zu negativen Einstellungen führen. Durch diese Interaktionen sollen Kinder lernen, Kulturen mit unterschiedlichen Regeln zu respektieren. Eine Evaluation zeigte jedoch, dass jüngeren Kindern häufig die emotionale Reife fehlt, um solche Konflikte konstruktiv zu verarbeiten. Daher ist eine altersgerechte Anpassung interkultureller Lernformate unerlässlich.



Abbildung 11: Adaptiertes Werwolf-Spiel, in denen das Kind mit virtuellen Gruppen konfrontiert wird, die nach anderen Regeln spielen

34 Ruth Aylett/Lynne Hall/Sarah Tazzyman/et al.: *Werewolves, cheats, and cultural sensitivity*, <https://dl.acm.org/doi/10.5555/2615731.2617418>



Abbildung 12: Rollenspiel mit einer virtuellen Job Interviewerin

In weiteren Forschungsprojekten haben wir Systeme zum emotionalen Lernen entwickelt, die Nutzerinnen und Nutzer auf emotional herausfordernde Situationen vorbereiten sollen. Im EU-Projekt TARDIS<sup>35</sup> interagieren beispielsweise Jugendliche in einem simulierten Bewerbungsgespräch mit einem virtuellen Charakter (siehe Abbildung 12). Im Rahmen dieses Projekts bzw. dem Nachfolgeprojekt EmpaT<sup>36</sup> wurde auch die zuvor diskutierte Simulation des Bewerbungsgesprächs als Vorstudie durchgeführt. Solche Gespräche sind für viele junge Menschen mit intensiven negativen Emotionen wie Nervosität, Stress oder Unsicherheit verbunden. Ziel des Trainings ist es, durch Rollenspiel mit der virtuellen Agentin den Umgang mit diesen Emotionen zu üben. Während der Interaktion werden Körpersignale der Teilnehmenden erfasst, um Rückschlüsse auf ihren emotionalen Zustand zu ermöglichen und

---

35 Tobias Baur/Ionut Damian/Patrick Gebhard/et al.: A Job Interview Simulation: Social Cue-Based Interaction with a Virtual Character, <http://dx.doi.org/10.1109/SocialCom.2013.39>

36 Patrick Gebhard/Tanja Schneeberger/Elisabeth André/et al.: Serious Games for Training Social Skills in Job Interviews, <http://dx.doi.org/10.1109/TG.2018.2808525>

das System entsprechend anzupassen. Eine Studie<sup>37</sup> zeigte, dass Jugendliche, die mit dem TARDIS-System trainiert hatten, in einem anschließenden Bewerbungsgespräch mit einem menschlichen Trainer signifikant besser abschnitten als eine Vergleichsgruppe, die sich mithilfe eines Leitfadens vorbereitet hatte. Die mit dem System trainierten Jugendlichen wirkten u. a. weniger nervös und zeigten mehr Blickkontakt – ein Hinweis auf ein gesteigertes Selbstvertrauen im Umgang mit der herausfordernden Situation.

#### **4.2 Anwendungen in der Psychotherapie**

KI-basierte Verfahren zur Verhaltensanalyse bieten ein zukunftssträchtiges Potenzial für die Diagnostik und Therapie psychischer Störungen. Durch die automatisierte Erfassung und Auswertung multimodaler Verhaltensmerkmale können subtile Anzeichen psychischer Erkrankungen objektiv erfasst und über den Zeitverlauf hinweg nachvollzogen werden. Schauen wir uns einige konkrete Beispiele an.

Die meisten bisherigen Arbeiten zur automatisierten multimodalen Bewertung psychischer Erkrankungen konzentrierten sich auf Depressionen. Ein Ansatz zur objektiven Erfassung depressiver Symptome ist die Analyse von Körperbewegungen im Behandlungsverlauf. Untersuchungen von Joshi et al. (2013)<sup>38</sup> zeigen, dass Personen mit schwerer Depression zu deutlich reduzierter Bewegungsaktivität neigen. Im Verlauf einer erfolgreichen Behandlung nimmt diese Aktivität in der Regel zu. Solche Veränderungen lassen sich mithilfe spezieller Bewegungsanalysen erfassen, etwa durch die Auswertung der räumlichen Ausrichtung und des Bewegungsumfangs einzelner Körperteile relativ zum Rumpf. Eine erhöhte Bewegungsintensität steht dabei oft im Zusammenhang mit einer Verbesserung des psychischen Zustands, gemessen z. B. anhand der Hamilton Rating Scale for Depression.<sup>39</sup>

---

37 Ionut Damian/Tobias Baur/Birgit Lugin/et al.: *Games are Better than Books, in: Situ Comparison of an Interactive Job Interview Game with Conventional Training*, [http://dx.doi.org/10.1007/978-3-319-19773-9\\_9](http://dx.doi.org/10.1007/978-3-319-19773-9_9)

38 Jyoti Joshi/Abhinav Dhall/Roland Goecke/et al.: *Relative Body Parts Movement for Automatic Depression Analysis*, <http://dx.doi.org/10.1109/ACII.2013.87>

39 [https://flexikon.doccheck.com/de/Hamilton\\_Depression\\_Scale](https://flexikon.doccheck.com/de/Hamilton_Depression_Scale)

Auch die Analyse gesprochener Sprache liefert aufschlussreiche Hinweise auf depressive Zustände. So zeichnen sich Sprachaufnahmen depressiver Personen häufig durch verminderte Energie, eine reduzierte Tonhöhenvariation und eine insgesamt monotonere Sprechweise aus. Dies lässt sich anhand von Spektrogrammen (Cummins et al. 2015)<sup>40</sup> sichtbar machen, bei denen Sprache auf ihre zeitlichen und frequenzbasierten Eigenschaften hin analysiert wird (vgl. Abbildung 3). Dabei zeigen sich bei nicht-depressiver Sprache klarere und differenziertere Frequenzmuster, während bei stark depressiver Sprache ein vermehrtes Rauschen und weniger Struktur zu erkennen ist. Diese Merkmale können automatisiert erkannt und in diagnostische Verfahren integriert werden.

Die Mimik ist einer der aussagekräftigsten Indikatoren für den aktuellen emotionalen Zustand eines Menschen. Bestimmte Veränderungen in der Gesichtsmuskulatur – etwa an Augen, Augenbrauen oder Lippen – können auf depressive Symptome hinweisen und lassen sich mittlerweile zuverlässig mit alltäglichen Geräten wie Smartphone-Kameras erfassen.

Studien, unter anderem von Song et al. (2022)<sup>41</sup>, haben gezeigt, dass bei depressiven Personen bestimmte Facial Action Units (AUs, siehe Abschnitt 2.1) in charakteristischer Weise auftreten:

AU4, die mit dem Herunterziehen der Augenbrauen verbunden ist, wird bei depressiven Personen häufiger aktiviert. Zudem sind ihre Dauer und Intensität oft erhöht.

AU12, die das Anheben der Mundwinkel beim Lächeln beschreibt, tritt hingegen seltener auf.

AU15, die ein Herabhängen der Mundwinkel signalisiert, ist bei depressiven Personen meist kürzer ausgeprägt.

AU17, die das Anspannen des Kinns betrifft, wird länger und häufiger gezeigt.

---

40 Nicholas Cummins/Vidhyasaharan Sethu/Julien Epps/et al.: *Analysis of acoustic space variability in speech affected by depression*, <http://dx.doi.org/10.1016/j.specom.2015.09.003>

41 Siyang Song/Shashank Jaiswal/Linlin Shen/et al.: *Spectral Representation of Behaviour Primitives for Depression Analysis*, <http://dx.doi.org/10.1109/TAFPC.2020.2970712>

Auch Daten aus psychotherapeutischen Gesprächen liefern wertvolle Hinweise zur Einschätzung des psychischen Zustands einer Person. Ein besonders aufschlussreicher Aspekt ist die Synchronizität – also die zeitliche Abstimmung von Reaktionen zwischen Gesprächspartnern. Diese kann als ein zentraler Indikator für Rapport, also das wechselseitige Verstehen und die emotionale Verbindung, sowie für das Engagement im Gespräch gelten, siehe auch Delaherche et al. (2012)<sup>42</sup> für einen Überblick.

Darüber hinaus eröffnet der Einsatz sensorbestückter Mobilgeräte neue Möglichkeiten, relevante Verhaltensdaten auch im Alltag außerhalb der Therapiesituation zu erfassen. Mit Smartphones oder sogenannten Wearables lassen sich beispielsweise Sprachverhalten, Bewegungsmuster, Herzfrequenzvariabilität oder Gesichtsausdrücke kontinuierlich und unaufdringlich aufzeichnen – ein vielversprechender Ansatz für eine alltagsnahe Bewertung psychischer Gesundheit, siehe das Übersichtspapier von Han et al. (2021)<sup>43</sup>.

Um psychotherapeutische Prozesse gezielt zu unterstützen, wurde von uns NOVA<sup>44</sup> entwickelt – ein (semi-)automatisches Annotations-tool zur Analyse verbaler und nonverbaler Verhaltensweisen (siehe Abbildung 13). NOVA macht aufgrund eines vortrainierten KI-Modells Vorschläge, die Psychotherapeuten und -therapeutinnen gegebenenfalls korrigieren können, wodurch das KI-Modell sukzessiv verfeinert wird.<sup>45</sup> In Zusammenarbeit mit Kolleginnen und Kollegen aus der Psychothe-

---

42 Emilie Delaherche/Mohamed Chetouani/Ammar Mahdhaoui/et al.: *Interpersonal Synchrony: A Survey of Evaluation Methods across Disciplines*, <http://dx.doi.org/10.1109/T-AFFC.2012.12>

43 Jing Han/Zixing Zhang/Cecilia Mascolo/et al.: *Deep Learning for Mobile Mental Health: Challenges and recent advances*, <http://dx.doi.org/10.1109/MSP.2021.3099293>

44 Alexander Heimerl/Katharina Weitz/Tobias Baur/Elisabeth André: *Unraveling ML Models of Emotion With NOVA: Multi-Level Explainable AI for Non-Experts*, <https://doi.org/10.1109/T-AFFC.2020.3043603>

45 Tobias Baur/Alexander Heimerl/Florian Lingenfeller/et al.: *eXplainable Cooperative Machine Learning with NOVA*, <https://doi.org/10.1007/s13218-020-00632-3>

rapie<sup>46</sup> zeigte sich, dass NOVA hilfreich ist, um mehrdeutige Gesprächsausschnitte mit geringer Übereinstimmung zwischen verbalen und non-verbalen Signalen zu identifizieren. Auch der kooperative Workflow für mehrere Annotierende wurde positiv bewertet, während der anfängliche Annotationsaufwand weiterhin als hoch eingeschätzt wurde.

Automatisierte Annotationswerkzeuge wie NOVA bieten neue Einblicke in emotionale Prozesse der Psychotherapie, die über klassische Fragebögen hinausgehen. So zeigte sich in einer weiteren Studie<sup>47</sup>, dass die automatisch erfassten Emotionsdimensionen von Erregung und Valenz signifikant mit den Einschätzungen der Therapeutinnen und Therapeuten zu Emotionen wie Traurigkeit, Angst oder Entspannung korrelierten, allerdings nicht mit den Selbstauskünften der Patientinnen und Patienten. Dies deutet darauf hin, dass maschinelle Analysen emotionale Zustände eher aus externer Beobachtungsperspektive erfassen. Die Maschine zieht im Wesentlichen die emotionalen Zustände von Patientinnen und Patienten heran, die direkt wahrnehmbar sind. Sie kann somit nicht in den Menschen „hineinschauen“. Als Ergänzung zur therapeutischen Arbeit kann NOVA dabei helfen, kritische Therapieverläufe frühzeitig zu erkennen.

Aufgrund der hohen Zahl kognitiv eingeschränkter Patientinnen und Patienten mit begrenzter sprachlicher Kommunikationsfähigkeit ist ein kontinuierliches Schmerz-Monitoring durch Pflegepersonal oder Angehörige langfristig kaum realisierbar. Maschinelle Lernverfahren zur automatischen Erkennung von Schmerzintensität und -qualität – vor allem anhand von Gesichtsausdrücken – bieten hier eine aussichtsreiche Ergänzung, insbesondere in klinischen und pflegerischen Kontexten.<sup>48</sup> Sie tragen u.a. dazu bei, Überdosierungen von Schmerzmitteln zu vermeiden. Allerdings steht für die Schmerzerkennung bislang nur eine be-

---

46 Tobias Baur/Sina Clausen/Alexander Heimerl/et al.: NOVA: A Tool for Explanatory Multimodal Behavior Analysis and Its Application to Psychotherapy, [http://dx.doi.org/10.1007/978-3-030-37734-2\\_47](http://dx.doi.org/10.1007/978-3-030-37734-2_47)

47 Patrick Terhürne/Brian Schwartz/Tobias Baur/et al.: Validation and application of the Non-Verbal Behavior Analyzer: An automated tool to assess non-verbal emotional expressions in psychotherapy, <https://doi.org/10.3389/fpsy.2022.1026015>

48 Elisabeth André/Miriam Kunz: Digitale Gesichts- bzw. Schmerzerkennung und ihr Potential für die klinische Praxis, in: Digitalisierung und Gesundheit, hg. von Alexandra Manzei-Gorsky/et al., Baden-Baden 2022, S.97–112.

grenzte Menge an Daten zur Verfügung. Aktuelle Forschungsarbeiten untersuchen deshalb, inwiefern sich Modelle zur Emotionserkennung für die Schmerzerkennung adaptieren lassen, etwa durch Fine-Tuning von auf Emotionserkennung trainierten neuronalen Netzwerken.<sup>49</sup>

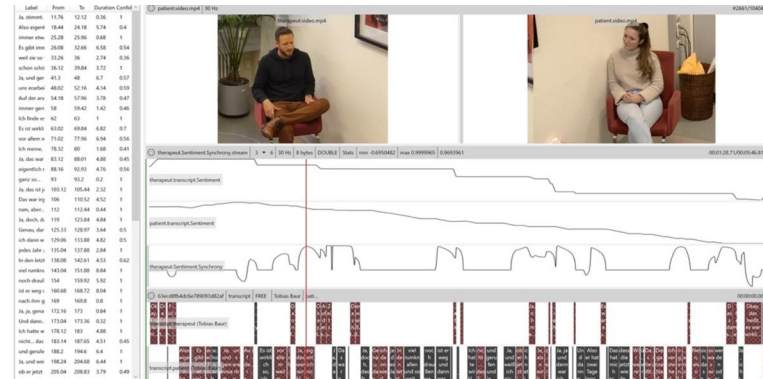


Abbildung 13: Screenshot einer mit dem NOVA System (semi-)automatisch annotierte nachgestellte psychotherapeutische Sitzung

## 5. Fazit

Ich fasse abschließend zusammen: Simuliertes emotionales Verhalten trägt nachweislich zur höheren Akzeptanz von humanoiden Robotern im Alltag bei. Besonders in sozialen Kontexten wirkt ein emotionsloser Roboter mit äußerlichen anthropomorphen Merkmalen auf viele Menschen befremdlich und unnatürlich. Emotionale Ausdrucksfähigkeit hilft hingegen, Erwartungen zu erfüllen, Vertrauen aufzubauen und Missverständnisse zu vermeiden.

Studien mit virtuellen Charakteren zeigen, dass diese großes Potenzial als Plattformen für emotionales und soziales Lernen bieten. Durch Rollenspiel mit virtuellen Charakteren können sich Personen in einer si-

49 Pooja Prajod/Dominik Schiller/Tobias Huber/Elisabeth André: Do Deep Neural Networks Forget Facial Action Units? – Exploring the Effects of Transfer Learning in Health Related Facial Expression Recognition, in: AI for disease surveillance and pandemic intelligence: intelligent disease detection in action, hg. von Arash Shaban-Nejad/et al., Cham 2022, S. 217–233.

cheren, kontrollierten Umgebung auf sozial herausfordernde Situationen vorbereiten, etwa auf Bewerbungsgespräche oder Konfliktsituationen.

In der Psychotherapie stößt der Einsatz emotionssensitiver Technologien zunehmend auf Interesse. Systeme, die alltägliche emotionale Zustände zu erfassen vermögen, helfen Therapeutinnen und Therapeuten dabei, Veränderungen im emotionalen Erleben von Patientinnen und Patienten zu erkennen.

Eine zentrale Herausforderung bleibt jedoch bei allen Anwendungen die fehlende Kontextsensitivität: Maschinen können zwar Emotionen simulieren und erfassen, jedoch noch nicht hinreichend deren Bedeutung im sozialen oder individuellen Zusammenhang vollständig verstehen. Darüber hinaus fehlt ihnen jegliche eigene körperliche Erfahrung.

Zu beachten ist auch, dass die Einführung empathischer Maschinen auch zu unerwünschten Effekten führen, siehe auch André (2014)<sup>50</sup> zur (teils ungesunden) Wirkung empathischer Maschinen und den daraus resultierenden ethischen Fragestellungen André (2015)<sup>51</sup>.

Damit solche Anwendungen nachhaltig und verantwortungsvoll eingesetzt werden können, bedarf es einer engen interdisziplinären Zusammenarbeit. Nur im Dialog zwischen Psychologie, Soziologie, Informatik und weiteren Disziplinen kann gewährleistet werden, dass emotionale Systeme sinnvoll in konkrete Anwendungsfelder integriert und zugleich kritisch begleitet werden.<sup>52</sup>

---

50 Elisabeth André: *Lässt sich Empathie simulieren? Ansätze zur Erkennung und Generierung empathischer Reaktionen anhand von Computermodellen*, Nova Acta Leopoldina NF 120 (2014), S. 81–105.

51 Elisabeth André: *Empathische Reaktionen und ihre Modellierung im Computer*, in: *Das soziale Gehirn. Neurowissenschaft und menschliche Bindung*, hg. von Helmut Fink/Rainer Rosenzweig, Paderborn 2015, S. 55–70.

52 Ich danke Andreas Eder für die Anfertigung der Transkription, die als Grundlage für die Ausarbeitung des vorliegenden Textes diente.

*Sven Nyholm*

## ETHIK DER KI: EINE PHILOSOPHISCHE PERSPEKTIVE

### *1. Der Bedarf ethischer Reflexion*

In diesem Beitrag möchte ich das Thema „Ethik der KI“ aus philosophischer Perspektive beleuchten. Dass das gesellschaftliche Großthema „KI“ bzw. „Künstliche Intelligenz“ philosophisch interessant ist und einer ethischen Analyse bedarf, werde ich Ihnen zu Beginn an einigen Beispielen veranschaulichen.

Wenn von Fortschritt in der Entwicklung von KI die Rede ist, dann wird häufig auf das Go-Spiel zwischen Weltmeister Lee Sedol und der KI AlphaGo verwiesen. Es fand im März 2016 statt. AlphaGo wurde in Googles KI-Schmiede DeepMind auf der Basis maschinellen Lernens trainiert und schlug Lee Sedol 4:1. Lee Sedol war danach am Boden zerstört und meinte, er höre jetzt mit Go auf. Es habe keinen Sinn mehr, denn er könne nicht mehr der beste Spieler werden.

Mir geht es bei diesem Beispiel aber nicht allein um den Sieg von AlphaGo. Mehr noch interessiere ich mich für eine andere Person, nämlich einen Mitarbeiter von DeepMind, der die Spielzüge, die von AlphaGo empfohlen wurden, ausführte.

AlphaGo ist am Ende kein Go-spielender Roboter, sondern ein Programm, das Empfehlungen abgibt. Es wurde trainiert und gefüttert mit Daten von Tausenden von Go-Spielen, die Menschen gespielt hatten, und verfeinerte danach in Millionen von Go-Spielen gegen sich selbst die erworbenen Strategiemuster. Um reale Auswirkungen zu haben, benötigte Alpha Go einen Menschen in räumlicher Präsenz zu Lee Sedol. Denn Arme, Hände, Raumbewusstsein – kurz alles, was für das Setzen der Steine notwendig ist, fehlten dem Programm. Dazu war also ein Mensch notwendig. Dieser Herr, der die Spielzüge auf Basis der Empfehlungen von AlphaGo ausführte, ist damit so gesehen eine Art menschlicher Roboter. Denn er musste weder die von AlphaGo gesetzten Züge noch die hinter ihnen stehenden Muster verstehen. Er konnte es wahrscheinlich auch gar nicht, denn die Züge waren neuartig und selbst für erfahrene

Spieler unerwartet. Er musste sie lediglich ausführen. Insofern wird er – so könnte man sagen – von AlphaGo programmiert oder gesteuert.

Deshalb lässt sich fragen: Wer hat hier eigentlich die Spiele gewonnen? Dass Lee Sedol sie verloren hat, das steht außer Frage. Aber wem schreiben wir die Siege zu? Dem Herrn von Google DeepMind, der nichts von Go versteht bzw. verstehen muss? Ist er der neue Weltmeister? Oder dem Computerprogramm selbst? Aber steht hinter diesem Programm nicht ein Unternehmen und aufgrund der Trainingsdatensätze viele tausend menschliche Spieler, die zum Erfolg beigetragen haben? Immerhin hat AlphaGo auch noch gegen sich selbst gespielt. Ich breche hier ab: Sie sehen, unsere gängigen Zuschreibungspraktiken geraten hier schnell an ihre Grenzen. Das Einzige, was wir einigermaßen klar benennen können, ist, dass Lee Sedol im März 2016 verloren hat.

Zwei Jahre später, im März 2018, ereignete sich mein zweites Beispiel. Es geht um die erste Fußgängerin, die durch ein selbstfahrendes Auto getötet wurde. Zuvor waren zwar bereits Fahrerinnen und Fahrer von selbstfahrenden Autos tödlich verunglückt. Aber zu diesem Zeitpunkt verstarb eine eigentlich unbeteiligte Person. Ein experimentelles selbstfahrendes Auto von Uber erfasste einen Fahrradfahrerin, die die Straße überquerte. Die zur Sicherheit eingesetzte FahrerIn (die sog. „safety driver“) hatte kurz die Aufmerksamkeit von der Straße genommen – ihre Rolle bestand eigentlich darin, solche Unfälle zu verhindern. Im folgenden Rechtsstreit wurde sie nach einem langen Prozess verurteilt. Doch die Debatte hat sich fortgesetzt, weil sich fragen lässt, ob man Menschen derart lange Konzentration abverlangen kann. Auch hier also ist klar: es gab zwei Verlierer. Und auch hier lässt sich nicht so eindeutig, wie es der Rechtsprozess voraussetzt, beantworten, wer eigentlich verantwortlich war.

Einige Philosophinnen und Philosophen argumentieren, dass die SicherheitsfahrerIn keine Schuld trifft. Denn während wir uns beim normalen selbstständigen Fahren gut über längere Zeit konzentrieren können, ist das im Modus der Untätigkeit nicht in gleichem Maße der Fall. Über eine sehr lange Zeit in dieser Habachtstellung zu verharren, das ist eigentlich unmöglich, ohne zu ermüden. Trifft das zu, haben wir auch hier eine Verantwortungslücke.

Das Thema Verantwortungslücke führt mich zu meinem dritten Beispiel. Über Verantwortungszuschreibung diskutieren wir auch dann,

wenn es um autonome Waffensysteme -im Umgangston auch „Killer Robots“ genannt – geht. Eine gesellschaftliche Kampagne, die sich für das Verbot solcher Roboter einsetzt, beruft sich hauptsächlich darauf, dass ihre Einführung zu Verantwortungsdiffusion und Verantwortungslücken führe. Die Gefahren lauerten nicht nur in der Tödlichkeit dieser Waffen, sondern auch in der Tatsache, dass sie unsere etablierten Vorstellungen von Verantwortung im Krieg außer Kraft setzten.

Nehmen wir einmal an, so ein Roboter tötet Zivilisten. Leider geschieht das auch ohne solche Roboter oft genug. Doch in den bisherigen Fällen setzt dann im besten Fall eine Untersuchung ein, die eruiert, wer für diesen Verstoß gegen das Kriegsrecht verantwortlich ist. So gehen wir bislang jedenfalls vor, denn die Zuschreibung von Verantwortlichkeit folgt klaren Regeln. Doch wen sollte man für „Handlungen“ von Killerrobotern verantwortlich machen?

Die Rede von einer Verantwortungslücke stützt sich u.a. darauf, dass neuronale Netze nach Prinzip „Blackbox“ funktionieren. Wir geben ihnen einen recht klar umrissenen Input und erhalten nach ihren Prozessen ein Ergebnis: einen Output. Wie die künstlichen neuronalen Netze zu diesen Ergebnissen kommen, ist jedoch selbst für Expertinnen und Experten des machine learning zumeist nicht nachvollziehbar. Hier setzt die Initiative „Explainable AI“ an, die versucht, die Funktionsprinzipien der Technologie durchsichtig zu machen.

Aber gehen wir erst einmal weiter in der Reihe meiner Beispiele. Das nächste ereignete sich im Juni 2022. Die Washington Post veröffentlichte damals einen Artikel mit einem Interview mit einem Softwareingenieur namens Blake Lemoine von Google. Dieser war entlassen worden, nachdem er das „LaMDA“ Language Model von Google für ein personales Wesen gehalten hat. Er war angestellt, um Gespräche mit diesem Chatbot zu führen. Und im Lauf dieser Gespräche ist er zu der Auffassung gelangt, dass er diesen Chatbots Gefühle, Intelligenz, Bewusstsein, ja kurz Person-Sein und Rechte zuschreiben müsse. Denn der Bot hätte ihn beauftragt, ihm einen Anwalt zu suchen, weil Google seine Rechte nicht achte. Dafür musste der Ingenieur Blake Lemoine viel Kritik einstecken, obwohl er einige interessante Argumente für seine Ansicht vorlegte, dass der Chatbot Bewusstsein und Emotionen entwickelt hatte. Zwar ist es derzeit noch Konsens, ihn für verwirrt zu halten oder die Argumente abzuweisen. Doch ich sehe das etwas anders. Meine Prognose ist, dass zunehmend mehr Menschen mit solchen Positionen auftreten werden

und mittlerweile gibt es in der Tat auch weitere Leute (inklusive KI-Forscherinnen und Forscher), die die Idee von bewussten KI-Systemen ernst nehmen.

Aber weiter zum nächsten Beispiel. Jetzt schreiben wir November 2022. Zur allgemeinen Überraschung der breiteren Öffentlichkeit – für die Tech-Szene war es technologisch nichts Neues – präsentierte OpenAI ChatGPT. Neu war daran eigentlich nur die breite Zugänglichkeit des Chatbots. Jede Person mit Internetzugang und Handy konnte jetzt die Ergebnisse von großen Sprachmodellen nutzen. Innerhalb kürzester Zeit zählte der Dienst mehrere Millionen Nutzerinnen und Nutzer. Das löste eine allgemeine Panik aus. Besonders an den Universitäten setzten hektische Debatten ein. Wie gehen wir ab jetzt mit Prüfungsleistungen um? Wie können wir verhindern, dass KI für Plagiate genutzt wird? Wie schulen wir unsere Studierenden im Umgang mit Fehlinformationen, die ausgegeben werden?

Was aktuell diskutiert wird, sind im Grunde genommen vor allem Erweiterungsversionen von ChatGPT und anderen großen Sprachmodellen. Wenn man die ChatGPT zugrundeliegende Technologie nimmt, man nennt sie Foundation Model, und wenn man sie mit weiteren Daten und anderen Inputs für einen bestimmten Zweck trainiert, kann man eine Reihe weiterer Anwendungen generieren. Ein interessantes Beispiel sind KI-Imitationen besonderer Individuen.

Da gibt es beispielsweise eine KI, die Daniel Dennett, den bekannten amerikanischen Philosophen, auf eine beeindruckende Art simuliert. Seine gesammelten Werke dienen als Grundlage, um ihn digital zu imitieren. Und die Outputs dieser Bots klingen tatsächlich erstaunlich genau wie diese Person. Dennett-Experten und Expertinnen können Textausgaben dieser KI-Bots („DigiDan“) kaum von originalen Werken unterscheiden.<sup>1</sup> Stil und Inhalt ähneln sich zu sehr.

Ähnliches wurde für Deepak Chopra gemacht. Er ist ein sehr berühmter Mental Health- und Wellnesscoach. Da er nicht alle Leute behandeln kann, die bei ihm eine Sprechstunde erbitten, gibt es jetzt digitale Versionen von ihm (sein KI-Zwilling „Digital Deepak“). Die sprechen dann passenderweise auch noch verschiedene Sprachen. Auch einzelne

---

1 Vgl. Eric Schwitzgebel/David Schwitzgebel/Anna Strasser: *Creating a Large Language Model of a Philosopher*, <https://doi.org/10.1111/mila.12466>

Influencerinnen haben sich bereits ChatBots gebaut, um den Anfragen ihrer Fans nach privaten Nachrichten gerecht werden zu können.

Eine weitere Anwendung, die in den USA und China bereits etabliert ist, verarbeitet Datensätze verstorbener Verwandter. Diese lassen quasi die im Datensatz aufbewahrte Person wiederauferstehen und Menschen können weiter in Kontakt bleiben, sich im Leben beraten lassen oder einfach Erfahrungen teilen. Ich habe neulich einen Artikel zu diesem Thema mit einem Kollegen aus China geschrieben, der mir versichert, solche digitalen Trauerpraktiken seien dort populär.<sup>2</sup>

## **2. Traditionelle Theoriebildung gerät an ihre Grenzen**

Angesichts dieser verschiedenen Beispiele erstaunt es nicht, dass von vielen Seiten eine philosophische Einordnung und Auseinandersetzung mit KI gewünscht wird. Da denken wir schnell nur an Fragen von rechtlicher Regelung, moralischer Bewertung und Ähnliches. Mich interessiert aber auch die zugrundeliegende philosophische Frage: Was ist eigentlich eine Form von Intelligenz? Worum handelt es sich bei ihr im Vergleich zu einem Menschen oder einem Tier?

Diese Art von Fragen halte ich für spannend, teilweise weil hier unsere traditionellen Begriffe an ihre Grenzen gelangen. Unsere traditionellen ethischen Theorien und auch unsere allgemeinen moralischen Intuitionen stammen aus einer Zeit ohne KI. Deshalb müssen wir ggf. unsere ethischen Vorstellungen und unsere Grundbegriffe aktualisieren.

Philosophische Fragen über Ethik der Künstlichen Intelligenz lassen sich in drei Bereiche gliedern: Erstens geht es um Fragen, die durch solche KI veranlasst werden, die bereits gesellschaftlich weit verbreitet und im Einsatz ist. Zweitens geht es um Fragen, die mit verbreiteten Vorstellungen über KI und Missverständnissen von KI zu tun haben. Hier ließe sich das Beispiel von Google LaMDA einordnen und die Debatte über die Bewusstseinsfähigkeit von KI. Drittens geht es um Fragen, die KI aufwirft, die bislang nur im Modus von Gedankenexperimenten und Extrapolationen vorliegen. Ein wenig nach der Logik von: Wenn KI

---

2 Vgl. Steven Campbell/Pengbo Liu/Sven Nyholm: *Can Chatbots Preserve Our Relationships with the Dead?*, <http://dx.doi.org/10.1017/apa.2025.1>

irgendwann so und so ist, wie wäre das dann (zu bewerten)? Teilweise klingt das nach Science-Fiction. Doch manchmal ist eben die Science-Fiction von gestern die Realität von heute.

Mittlerweile gibt es im Übrigen keinen Mangel an ethischen Zugriffen auf das Thema. Dafür müssen wir nur einen Blick in zwei gängige philosophische Überblickswerke werfen, zum Beispiel in die Stanford Encyclopedia of Philosophy<sup>3</sup> oder die Internet Encyclopedia of Philosophy<sup>4</sup>. Die Inhaltsverzeichnisse der Hauptartikel zur Ethics of AI listen dort die gegenwärtigen Debatten auf: Es geht um Privacy und Surveillance, es geht um Bias und Explainability, es geht um Automatisierung und die Zukunft der Arbeit, es geht um Human-Robot Interaktion. Und so weiter und so fort.

Ich kann darum heute nur einen kleinen Einblick geben. Beginnen wir einmal mit der Farge: Wer forscht eigentlich im Bereich Ethik der KI?

Da finden wir drei Gruppen: Die erste wird gebildet durch Philosophinnen und Philosophen, wie z.B. die Humboldt-Professoren und Professorinnen Vincent Müller (FAU Erlangen) und Aimee van Wynsberghe (Uni Bonn). Darunter sind auch an der Philosophie interessierte Theologen und Theologinnen, wie beispielsweise Anna Puzio, eine begabte junge deutsch-polnische Theologin, die mehrere interessante Aufsätze und Bücher verfasst hat. Die zweite Gruppe besteht aus Informatikerinnen und Informatikern, darunter etwa Johanna Bryson und Virginia Dignum, die beide anfangs Informatikerinnen waren, nun aber einflussreiche Professorinnen im Bereich der Ethik der KI sind. Die dritte Gruppe besteht aus Menschen, die im Big Tech Bereich arbeiten. Alle großen Firmen von Google über Open AI bis SAP beschäftigen Ethikerinnen und Ethiker. Wir haben jedoch schon gesehen, dass so ein Job bei einem Tech-Giganten auch riskant werden kann. Denken wir an Timnit Gebru und Blake Lemoine, die bei Google über Ethik der KI geforscht haben, aber Google kritisiert haben und dann entlassen wurden.

Zwei der wichtigsten Forscherinnen und Forscher, die zur Zeit mit Ethik der KI bei Google arbeiten, sind Arianna Manzini und Iason Gabriel. Beide arbeiten bei Google beispielsweise an der ethischen Bewer-

---

3 <https://plato.stanford.edu/entries/ethics-ai/>

4 <https://iep.utm.edu/ethics-of-artificial-intelligence/>

tung dessen, was sie „advanced AI assistants“ nennen.<sup>5</sup> Aus deren Papers lässt sich oft herauslesen, was gerade die neuesten Entwicklungen bei Google und Co sind. Beispielsweise konnte man in ihren Texten schon lange bevor es allgemein bekannt war über die large language models lesen. The „next big thing“, um in der Sprache der Tech-Branche zu bleiben, dürften autonom agierende KI-Assistenten sein. Da wird dann auf ein LLM eine Art personales Overlay gelegt. Und diese Modelle können dann als private Angestellte arbeiten, wir könnten scherzhaft sagen, beinahe als KI-Butler oder Diener. Diese füttern wir mit Daten über unser Leben und sie erfüllen dann Alltagsaufgaben: Planen unseren Tag, buchen Zugtickets usw. Das ist auf jeden Fall eine der Visionen über die nächsten Schritte in der Entwicklung von KI bei Google DeepMind.

### **3. Der Begriff „Künstliche Intelligenz“**

Fokussieren wir nun den Begriff der Künstlichen Intelligenz. Zunächst eine kurze Begriffsgeschichte. Er taucht das erste Mal 1955 in einem Forschungsantrag auf. Aber ich würde sagen: die Idee ist alt. Sogar Aristoteles hat sich schon Gedanken gemacht über etwas Ähnliches. Er stellt ein Gedankenexperiment an, in dem er Werkzeugen Selbsttätigkeit zuschreibt und die daraus folgenden Konsequenzen für die Sozialstruktur beschreibt:

„Denn freilich, wenn jedes Werkzeug auf erhaltene Weisung, oder gar die Befehle im Voraus erratend, seine Verrichtung wahrnehmen könnte, wie das die Statuen des Dädalus oder die DreifüÙe des Hephästus getan haben sollen [...] wenn so auch das Weberschiff von selber webte und der Zitherschläger von selber spielte, dann brauchten allerdings die Meister keine Gesellen und die Herren keine Knechte.“<sup>6</sup>

Bei Aristoteles ist also schon der Gedanke einer Revolution der Arbeitswelt angelegt, die Aufhebung der Unterscheidung von Herr und Knecht. Solche selbsttätigen Werkzeuge würden gerade die Aufgaben

---

5 Vgl. Iason Gabriel/Arianna Manzini/et al.: *The Ethics of Advanced AI Assistants*, <https://doi.org/10.48550/arXiv.2404.16244>

6 Aristoteles: *Der Staat der Athener*, hrsg. u. übers. v. Martin Dreher, Stuttgart 1986, 1253b–1254a.

übernehmen, zu denen im alten Griechenland Menschen gezwungen wurden.

Aber zurück in die 1950er Jahre. Während es davor so etwas wie gedankliche Experimente gab, ändert sich hier folgendes: Man beginnt, auch die technische Seite mit in Angriff zu nehmen. Wie kommt es zu der Wortverbindung Künstliche Intelligenz? Ich denke, dahinter liegt folgende Beobachtung: Es handelt sich um Technologien, welche Aufgaben übernehmen, die normalerweise bzw. bisher menschliche Akteure mittels ihrer Intelligenz ausgeführt haben. Es sind Dinge wie z.B. Texte schreiben, Auto fahren, Empfehlungen machen, Entscheidungen treffen, medizinische Diagnosen stellen. Diese Tätigkeiten übernimmt nun eine „Maschine“ oder ein Computer, also eine Technologie, die in diesem Sinn künstlich intelligent ist.

Wir nennen sie deshalb intelligent, weil sie ähnliche Dinge bewirken können wie menschliche Intelligenz. Das wäre ungefähr der Kern von Alan Turings Gedanken. Anstatt zu fragen, ob diese Technologien selbst denken können, meinte Turing, dass wir uns darauf fokussieren sollten, ob sie uns Menschen imitieren können.<sup>7</sup> Und wenn sie es können, spräche nichts dagegen, ihnen künstliche Intelligenz zuzuschreiben. Maschinelle Intelligenz bestünde dann laut Turing darin, Verhalten imitieren zu können, das zuvor von menschlicher Intelligenz abhing.

In einem Forschungsvorschlag von John McCarthy und Kollegen aus dem Jahr 1955, der den Begriff „Künstliche Intelligenz“ in unsere Alltagssprache einführte, ging es um Maschinen, die Intelligenz simulieren können. Hier findet sich also ein feiner Unterschied zu Turings Formulierung. Seit den 1990ern werden KIs dann als „intelligente Agenten“ verstanden. Die einflussreichen Informatiker und KI-Forscher Stuart Russell & Peter Norvig sprechen davon, dass intelligente Maschinen solche sind, die ihre Umgebung wahrnehmen können und die aus solchen Wahrnehmungsdaten abgeleitet handeln, also Ziele verfolgen, können.<sup>8</sup> Das ist eine neuartige Definition. Die beiden Autoren definieren Intelligenz dabei so: Intelligent ist ein Agent genau dann, wenn er Ziele auf

---

7 Vgl. Alan Turing: *Computing Machinery and Intelligence*, *Mind* LIX 1950, S. 433–460.

8 Vgl. Stuart Russell/Peter Norvig: *Artificial Intelligence: A Modern Approach*, New York 2022.

eine effektive Art verfolgen kann. Das können laut Russell & Norvig sowohl Menschen als auch Maschinen sein. Beide fallen unter diese Begriffsdefinition.

Der italienische Philosoph Luciano Floridi hingegen ist an dieser Stelle anderer Meinung. Er behauptet, es geht hier nicht im eigentlichen Sinn um Intelligenz.<sup>9</sup> Vielmehr sei zentral, dass KI einen neuen Typ Akteur einführt. Diese neuartigen Akteure verändern ihre Umwelt und führen dazu, dass wir uns als Menschen mit jeder Menge neuer Sachverhalte auseinandersetzen müssen. Ich bin hier anderer Meinung. Ich denke sehr wohl, wir können über eine neue Art von Intelligenz reden, aber eben über zwei verschiedene Typen: menschliche und künstliche, die sich doch auch unterscheiden. Das verbindende Element zwischen den beiden Typen besteht darin, dass künstliche Intelligenz zunehmend Aufgaben übernimmt, die früher menschliche Intelligenz ausgeführt hat.

Gehen wir einen Schritt weiter und blicken wir darauf, wie die großen Tech-Firmen heute KI definieren. Hier geht es häufig um die Entwicklung von sog. Artificial General Intelligence (AGI), bzw. Allgemeine Künstliche Intelligenz. Dieser Begriff steht für die Zukunftsvision vieler der großen Unternehmen. Es geht darum, die Flexibilität und die Kreativität menschlicher Intelligenz in einer artifiziellen Technologie abzubilden. Das hieße dann auch, Moralität, Sittlichkeit, ja kurz das Humanum schlechthin in Technologie zu bergen. Diese Vision prägt sowohl OpenAI als auch Google DeepMind. Sie verbinden damit eine fundamentale Revolution unserer Lebensweise, weil solche AGI uns von allen Aufgaben, die wir ungern erledigen, entlasten soll und zudem alte Menschheitsprobleme angehen würde. Teilweise halten einzelne Beobachter die Schwelle zur AGI schon für überschritten, beispielsweise Blaise Agüera y Arcas und Peter Norvig. Sie argumentieren, dass große Sprachmodelle zwar halluzinieren und noch allerlei Arten von logischen Fehlern begehen, aber dennoch bereits eine erste Art von AGI darstellen.<sup>10</sup> Als Hauptargument dafür führen sie an, dass schon heutige LLMs mehrere Arten von Aufgaben ausführen können.

---

9 Vgl. Luciano Floridi: *The Ethics of Artificial Intelligence: Principles, Challenges, and Opportunities*, Oxford 2023.

10 Vgl. Blaise Agüera y Arcas/Peter Norvig: *Artificial General Intelligence is Already Here*, Noema Magazine 2023. <https://www.noemamag.com/artificial-general-intelligence-is-already-here/>

Aber stimmt das denn? Haben wir schon eine Art von AGI in großen Sprachmodellen? Ich wäre da skeptischer. Immerhin scheint die semantische Treffsicherheit der Modelle eher begrenzt zu sein. Ein Beispiel wäre der Bild-Prompt „Lachs im Fluss“, der fein filetierte Lachsstücke schwimmend in einem Fluss ausgibt.<sup>11</sup> Das ist ohne Frage witzig und auch spannend. Aber es lässt sich schon fragen, wie allgemein intelligent ein solches Modell tatsächlich ist.

#### 4. *Ethische Herausforderungen der KI*

Grundsätzlich lässt sich unterscheiden zwischen einem engen und einem weiten Verständnis von Ethik der KI. Unter „eng“ verstehe ich die Debatte, die sich damit beschäftigt, welche konkreten Technologien wir tolerieren sollten und welche nicht. In einem breiten Verständnis hingegen diskutieren wir eher die Frage, welche Rolle diese Technologien in unseren Leben spielen sollten. Hier fokussieren wir eher die Rückwirkungen der Technologie auf die menschliche Lebensform und ihre einzelnen Handlungszusammenhänge.

Dem entspricht zum Teil die Unterscheidung zwischen negativer und positiver Ethik der KI. Während erstere sich darauf fokussiert, was verhindert werden muss, fragt positive Ethik der KI danach, welche Zukunftsvisionen sich durch die Integration von KI in unser Leben ergeben.

Drittens ist die bloße Idee von KI schon an sich ethisch interessant. Warum? Weil sie einen neuen Typ von Akteur einführt. Nämlich eine Technologie, die Aufgaben übernimmt, die normalerweise menschliche Intelligenz übernommen hat. Solche Aufgaben sind im Regelfall aber Aufgaben, die für Menschen wichtig sind. Es lässt sich deshalb fragen:

1. Verlieren wir Fähigkeiten, weil wir sie nicht mehr trainieren, wenn wir sie an digitale Technologie abgeben?
2. Gibt es noch ausreichend erfüllende Aufgaben für Menschen, wenn KI sie übernehmen kann?

---

<sup>11</sup> <https://www.globalnerdy.com/2024/07/14/when-ai-is-asked-to-make-a-picture-of-salmon-in-a-river/>

3. Sollen KI-Systeme Entscheidungen für uns treffen, die wir früher selbst getroffen haben?
4. Gibt es Entscheidungen, die generell nicht von KI getroffen werden sollten?

Zum Punkt 1 ließe sich beispielsweise fragen, ob Studierende noch ausreichend Grundfähigkeiten erwerben, wenn sie Hausarbeiten nicht mehr selbst schreiben. Denn eine Hausarbeit zu schreiben trainiert ja wichtige Grundfähigkeiten, es geht weniger um das fertige Ergebnis.

Wir dürfen in diesem Zusammenhang auch nicht vergessen, dass wir manche Aufgaben auch mit einem inneren Wert versehen. Sie auszuführen ist dann an sich gut. So wäre eine Zukunft, in der Menschen die einfachen Jobs erledigen, damit die KI Gedichte schreibt und Kunstwerke malt, keine erstrebenswerte Zukunft. Das veranschaulicht Punkt 2.

Unter Punkt 4 kann man die Diskussionen einordnen, ob KI Entscheidungen über Leben und Tod treffen sollten. Ein klassisches Beispiel wäre das bekannte sog. Trolley-Problem, aber nicht mehr mit einem Menschen als Entscheider, sondern einem selbstfahrenden Auto. Es gibt Ethikerinnen und Ethiker, die es per se für problematisch halten, wenn wir solche Entscheidungen an KI delegieren. Andere widersprechen und weisen darauf hin, dass in manchen Situationen eben entschieden werden muss und es dann immerhin besser wäre, wenn KI ethischen Regeln folge. Sie sollen also regelbasiert entscheiden und nicht willkürlich.

Hier stellt sich die Anschlussfrage: kann eine KI, sofern sie kein Mensch ist, überhaupt eine moralische Entscheidung treffen? Oder anders formuliert: Ist eine KI eigentlich ein (vollwertiger) moralischer Akteur? Hier gehen die Ansichten auseinander. Mark Coeckelbergh beispielsweise meint, um ein moralischer Akteur zu sein, müsse man Emotionen, moralische Emotionen haben.<sup>12</sup> Carissa Véliz geht noch einen Schritt weiter und behauptet, dass zusätzlich auch noch Bewusstsein vorliegen müsse. Und zwar als notwendige Bedingung: ohne Bewusstsein kann es keinen moralischen Akteur geben.<sup>13</sup>

---

12 Vgl. Mark Coeckelbergh: *Moral Appearances: Emotions, Robots, and Human Morality*, <http://dx.doi.org/10.1007/s10676-010-9221-y>

13 Vgl. Carissa Véliz: *Moral Zombies: Why Algorithms Are Not Moral Agents*, <https://link.springer.com/article/10.1007/s00146-021-01189-x>

Daraus ergibt sich die Gegenfrage: Ist es unmöglich, bewusste KI-Systeme zu entwickeln, die zusätzlich emotionale Einstellungen entwickeln? Auch hier gehen die Ansichten auseinander: vom oben genannten Google-Ingenieur Blake Lemoine, der entlassen wurde, über Ansichten von moderatem Bewusstsein schon bei heutigen KIs bis hin zu Positionen, die es für a priori ausgeschlossen halten, dass Systeme emotionale Einstellungen entwickeln können. Hier geht es vor allem darum, was wir sog. neuronalen Netzwerken zutrauen. Neben Elon Musk hält auch der Nobelpreisträger Geoffrey Hinton Chatbots bereits für sentient beings. Hinton versteht darunter Wesen, die sowohl ihre Umgebung wahrnehmen können als auch daraus abgeleitet spezifische Outputs produzieren können. Das könne GTP 4.0. Natürlich kann man streiten, ob damit schon alles abgedeckt ist, was wir gemeinhin unter Bewusstsein verstehen. Aber Hinton's Beitrag zeigt, dass hier Bewegung in der Debatte ist. Die großen Techfirmen sind jedenfalls überzeugt davon. Beispielsweise hat Anthropic einen „AI welfare“ Forscher eingestellt, der eruieren soll, ob Menschen zukünftigen KI-Systemen gegenüber moralische Pflichten besitzen, wie auch immer mit Bewusstsein begabt diese KI-Systeme sind.

Das legt den Gedanken nahe, dass es sich hier um neue moralische Subjekte und Objekte handeln könnte. Moralische Subjekte sind Wesen, die moralische Entscheidungen treffen können und ihr Handeln an solchen Entscheidungen orientieren können: Menschen beispielsweise. Moralische Objekte sind solche Wesen, denen gegenüber moralische Subjekte Pflichten haben können: zum Beispiel Tiere. Diese beiden Eigenschaften lassen sich kreuzen. Ein (erwachsener) Mensch ist sowohl ein moralisches Subjekt als auch ein moralisches Objekt. Ein Tier ist ein moralisches Objekt, im Regelfall aber kein moralisches Subjekt. Ein Stein ist weder moralisches Subjekt noch Objekt. Wo in dieser Aufteilung taucht nun aber die KI auf?

Gibt es möglicherweise Entitäten, die zwar eine Art moralische Subjekte sind, wie selbstfahrende Autos, aber keine moralischen Objekte sind? Oder sind KI-Systeme wie Steine: weder moralische Subjekte noch moralische Objekte?

Ethikerinnen und Ethiker ordnen hier sehr unterschiedlich zu. Sollten wir es aber mit einem neuen ethischen Akteur zu tun haben, dann brauchen wir wohl auch neue ethische Prinzipien. Bislang haben wir vor allem Prinzipien für die Mensch-Mensch Interaktion. Wir bräuchten

dann auch solche für die Mensch-KI-Interaktion und für die KI-KI-Interaktion.

Nehmen wir einmal das Beispiel der von Google vorgeschlagenen „Advanced AI Assistants“. Sie würden Termine zwischen Menschen vereinbaren. Dafür müssen sie sowohl untereinander (KI-KI) als auch mit ihren jeweils zugeordneten Menschen interagieren (KI-Mensch). Die Frage wäre dann: Sollen für die Interaktion zwischen KI und KI dieselben moralischen Regeln gelten wie von Mensch zu Mensch?

Nehmen wir einmal die Frage nach der Erklärbarkeit. Manche Ethikerinnen und Ethiker vertreten, dass wir von KI, weil sie mehr kann als Menschen, mehr Erklärbarkeit erwarten sollten, also von einem selbstfahrenden Auto einen höheren Grad der Sicherheit als von einem „bloßen“ Menschen.

Oder denken wir an den Fall eines Chatbots von Air Canada. Ein Chatbot hatte einem Kunden eine falsche Information über einen speziellen Typ Flug gegeben. Air Canada hatte dann behauptet, dass für den Fehler allein der Chatbot verantwortlich wäre, nicht aber das Unternehmen. Hier sehen wir also, dass Unternehmen vermeintliche Verantwortungslücken versuchen auszunutzen.

Um hier klarer zu sehen, müssen wir den Begriff Verantwortung weiter unterscheiden. Und zwar anhand der beiden Achsen positiv-negativ und retrospektiv-prospektiv. Negative Verantwortung hat mit Tadel und Bestrafung zu tun, positive mit Handlungen, auf die wir mit Anerkennung und Lob reagieren. Zusätzlich können wir entweder für vergangene Handlungen verantwortlich sein oder für zukünftige Folgen unseres Handelns. Innerhalb dieser vier Achsen gibt es dann verschiedene potentielle Typen von Verantwortungslücken.

Im Fall der Übergabe von Aufgaben an KI-Systeme lässt sich dann z.B. fragen: Wer verdient Lob, wenn die Aufgaben gut erfüllt werden und wer verdient Tadel, wenn etwas schiefgeht?

## **5. Zusammenfassung**

Ich fasse zusammen: Ich habe ein breites Feld skizziert, wo und wie KI-Systeme bisher dem Menschen vorbehaltenen Aufgaben übernehmen. Aus diesem Faktum folgt: Unsere traditionellen ethischen Theorien müs-

sen für das Zeitalter der KI aktualisiert werden. Schon die Ontologie von KI an sich ist interessant. Stellen sie moralische Subjekte, Objekte oder beides zugleich dar? Ausgehend davon wären Prinzipien zu etablieren für drei Typen von Interaktion: Mensch-Mensch; Mensch-KI und KI-KI. Damit sind die Aufgaben der kommenden Jahre dargelegt.<sup>14</sup>

---

14 Für seine Hilfe, die schriftliche Version dieses Textes zu erstellen, danke ich Andreas Eder. Dieser kurze Text basiert auf einem in Kürze erscheinenden Buch, und zwar Sven Nyholm, *The Ethics of Artificial Intelligence. A Philosophical Introduction*, das bei Hackett Publishing (Cambridge Mass.) 2025 erscheinen wird.

*Christian Albrecht*

## NATÜRLICHE INTELLIGENZ. THEOLOGISCH-ETHISCHE ASPEKTE DES EINSATZES VON KI IM CHRISTLICHEN SOZIAL- UND GESUNDHEITSWESEN

Alles Gute kommt aus Amerika, und es ist dort früher entstanden als anderswo. So ist es auch mit der Nutzung von KI im Sozial- und Gesundheitswesen. Viele positive Erfahrungen hat man dort gemacht, zum Beispiel beim Aufnahme- und Entlassungsmanagement in Krankenhäusern durch einen Chatbot, weil der „alle Fragen ‚mit endloser Geduld‘ auch doppelt und dreifach“ beantwortet. Nur ab und zu passieren dem Chatbot dabei auch Fehler, zum Beispiel, dass er einem Patienten abschließend die Scheidung empfiehlt.<sup>1</sup>

KI-Systeme haben, es kann nicht oft genug daran erinnert werden, kein Bewusstsein und sie haben auch kein Verständnis, sondern, wenn überhaupt, nur eine maximal eingeschränkte, automatisierte Form des Verständnisses. Künstliche Intelligenz ist nicht intelligent, sondern kann nur intelligentes Verhalten simulieren.

Doch gerade in dieser unendlich ausdehnungsfähigen Simulation von Intelligenz liegen bekanntlich die enormen Chancen von KI, die Chancen zur Vereinfachung und Verbesserung von Vorgängen in vielen Bereichen wirtschaftlicher Unternehmen und öffentlicher Verwaltung, in vielen Bereichen der Medizin und Bildung, in vielen Bereichen des alltäglichen Lebens und so auch im Bereich des christlichen Sozial- und Gesundheitswesens.

Auch wenn dieser Einsatz von KI im christlichen Sozial- und Gesundheitswesen ein weniger komplexer Fall ist als etwa die Problematik von KI in autonomen Autos oder in autonomen Waffensystemen, gilt auch hier: Der Einsatz von KI muss gestaltet werden, damit er menschendienlich bleibt. Aber was heißt das? Was ist ein menschendienli-

---

<sup>1</sup> So Alena Buyx im Vortrag zum Start der Fachmesse ConSozial am 16. Oktober 2024 in Nürnberg, vgl. <https://www.altenheim.net/medizinetheke-rin-warnt-vor-blindem-vertrauen-in-ki/>

cher Einsatz von KI, was ist ein menschendienlicher Einsatz von KI aus theologisch-ethischer Perspektive?

In der theologisch-ethischen Beurteilung von KI dominiert der kritische Blick auf Heilsversprechen, die sich mit KI verbinden und die Erinnerung an realistische Grenzen ihres Einsatzes, ihres aktuellen und potentiellen Einsatzes.<sup>2</sup> Das scheint mir in einem grundsätzlichen Blick auf KI zwar nachvollziehbar, trifft aber die speziellen Aufgaben einer theologisch-ethischen Reflexion auf Recht und Grenzen der KI im christlichen Gesundheits- und Sozialwesen nicht ganz. Einerseits sind hier die Heilerwartungen von Haus aus kleiner formatiert, andererseits sind die Untergangssängste nicht so ausgeprägt wie in anderen Bereichen. In theologisch-ethischer Perspektive geht es hier viel mehr um die Frage nach Kriterien des legitimen Einsatzes von KI. Welche Kriterien gibt es, das Unterstützungspotential von KI in der christlichen Sozialarbeit zu nutzen oder abzulehnen? Diese Unterstützungspotentiale von KI sind bekanntlich groß, sie sind noch längst nicht ausgeschöpft und sie werden immer größer. Genau deswegen muss der Einsatz von KI in der christlichen Sozialarbeit bewusst gestaltet werden, man kann nicht einfach warten, was kommt und was funktioniert. Was will man, was will man nicht? Welche Kriterien gibt es, die begründete Entscheidungen erlauben darüber, welche Aufgaben man KI überträgt, bei welchen Aufgaben man sich von KI unterstützen lässt und bei welchen Aufgaben man bewusst auf mögliche Unterstützungen durch KI verzichtet? Im theologischen Nachdenken über KI in der christlichen Sozialarbeit geht es also meines Erachtens weniger um die vielbeschworenen apokalyptischen Visionen der Selbstunterwerfung des Menschen unter die Maschine, die ihn in eine Abhängigkeit führe, aus der er nicht mehr herauskomme. Es geht im theologischen Nachdenken vielmehr um die Frage, welche konkreten Zuständigkeiten man lernenden Systemen übertragen möchte, genauer: um die Frage nach den Kriterien für diese Übertragung: Was geht, was geht nicht? Kriterien hierfür findet man jedenfalls nicht im Bereich apokalyptischer Schreckensvisionen oder im Zusammenhang eschatologischer Allversöhnungsidyllen, die von der Theologie hier bisweilen aufgerufen werden. Das hilft nicht weiter. Ich meine hingegen, man kann solche Kriterien in der Schöpfungslehre finden.

---

2 *Yannick Schlote: Konvergenz und Überwältigung. Die Mythen der Künstlichen Intelligenz aus theologisch-ethischer Perspektive, München/Hildesheim 2023.*

Und damit ist bereits gesagt, wie ich das Thema bearbeiten möchte. In einem ersten Abschnitt möchte ich mich solchen Kriterien in der Schöpfungslehre zuwenden, genauer gesagt: Ich möchte drei Kriterien nennen, die ich aus der Schöpfungslehre herausarbeiten werde und die eine Art Leitfaden für den Umgang mit KI in der christlichen Sozialarbeit geben können. In einem zweiten Abschnitt möchte ich die Anwendung dieser Kriterien konkretisieren, genauer gesagt: Ich möchte die genannten Kriterien auf verschiedene Anwendungsbereiche von KI im Gesundheits- und Sozialwesen beziehen und dabei drei Typen von Anwendungen identifizieren: erstens unproblematische Anwendungen, zweitens problematische, aber unter bestimmten Bedingungen denkbare Anwendungen und drittens ausgeschlossene, abzulehnende Anwendungen.

## 1. Kriterien der Schöpfungslehre

Für das evangelische Verständnis der Schöpfung ist Luthers Auslegung des ersten Glaubensartikels in den Katechismen zentral. Im Kleinen Katechismus beginnt die Auslegung bekanntlich mit dem Satz: „Ich glaube, dass mich Gott geschaffen hat samt allen Kreaturen“<sup>3</sup>. Das heißt: der Mensch versteht sich selbst als Geschöpf, das in doppelter Beziehung steht: zum Schöpfer hin, „der mich geschaffen hat“ und zu allem anderen hin, was Gott geschaffen hat „samt allen Kreaturen“<sup>4</sup>. Das heißt: „Der Mensch ist nach Leib und Seele Geschöpf“ und „existiert als Geschöpf in einer Welt, die Gott geschaffen hat.“<sup>5</sup> Weiter geht es im Kleinen Katechismus: Ich glaube, dass Gott mir „Leib und Seele, Augen, Ohren und alle Glieder, Vernunft und alle Sinne gegeben hat und noch erhält“<sup>6</sup>. Gottes

---

3 *Die Bekenntnisschriften der evangelisch-lutherischen Kirche, hrsg. im Gedenkjahr der Augsburgerischen Konfession 1930, 12. Auflage Göttingen 1998, S. 510, Z. 33f.*

4 *Evangelischer Erwachsenenkatechismus. Kursbuch des Glaubens. Im Auftrag der Katechismuskommission der Vereinigten Evangelisch-Lutherischen Kirche Deutschlands herausgegeben von Werner Jentsch, Hartmut Jetter, Manfred Kießig und Horst Reller, Gütersloh 31977, S. 179.*

5 *Martin Honecker: Art. Schöpfung IX. Ethisch, in: TRE 30 (1999). S. 348–355, 348.*

6 *BSLK (s.o. Anm. 3) S. 510, Z. 34–36.*

Schöpfertätigkeit ist also kein einmaliger Akt, sondern ein dauerhafter Vorgang, in dem Gott dank seiner Fürsorge die Welt erhält. Und schließlich wird im Kleinen Katechismus auch angesprochen, was daraus für den Menschen folgt: er hat Gott „zu danken und zu loben und dafür zu dienen und gehorsam zu sein“<sup>7</sup>. Der Mensch hat also seiner Geschöpflichkeit entsprechend zu handeln. Der christliche Schöpfungsgedanke, so bringt Luther hier in der denkbar komprimiertesten Weise zum Ausdruck, enthält also drei Aspekte: erstens bestimmt er das Bewusstsein des Menschen als eines Wesens, das sich und sein Leben Gott verdankt; zweitens unterstreicht er, dass der Mensch in Beziehung steht zu einer ebenfalls als geschöpflich verstandenen Welt und drittens hebt er hervor, dass der Mensch in dieser Welt etwas tun soll und wie er es tun soll, nämlich: so zu handeln, dass es dieser Geschöpflichkeit entspricht.

Luther bringt hier Aspekte zum Ausdruck, die teils schon in der Theologie vor Luther, vor allem aber in der evangelischen Theologie nach Luther zentral geworden sind in den vielen und ausführlichen Ausarbeitungen des Lehrstückes von der Schöpfung. Das werde ich hier keinesfalls auch nur ansatzweise referieren. Ich möchte aber drei Aspekte herausgreifen und unterstreichen, die in vielen dogmatischen und ethischen Ausarbeitungen anklingen und die mir für eine Orientierung in unserer Fragestellung, für die Suche nach Kriterien des Umgangs mit KI, von großer Bedeutung zu sein scheinen.

### *1.1 Die Offenheit der Schöpfung für menschliches Handeln*

Der erste Aspekt betrifft die Offenheit der göttlichen Schöpfung für menschliches Handeln. Dieser Aspekt ist angelegt in dem Gedanken, dass der von Gott geschaffene Mensch in der von Gott geschaffenen und auf den Menschen hin geordneten Welt aktiv und im Sinne Gottes handeln soll. Das muss erklärt werden. Gottes Schöpfung ist nicht statisch, irgendwann einmal gemacht und vollendet. Vielmehr ist Gottes Schöpfungsakt so angelegt, dass er nach einer Fortsetzung in einer gemeinsamen Geschichte von Schöpfer und Geschöpf verlangt. Die Schöpfung ist nicht abgeschlossen, sondern offen für die Zukunft, und zwar für eine vom Menschen zu gestaltende Zukunft. In der dogmatischen Tradition wird das teils aufgenommen in dem engen Zusammenhang

---

7 BSLK (s.o. Anm. 3) S. 511, Z. 6f.

von Schöpfung und Erhaltung: Gott schafft und erhält sein Werk. Vor allem aber wird das aufgenommen in der vielfach missverstandenen Aufforderung Gottes an den Menschen, sich die Erde untertan zu machen. Recht verstanden, bedeutet diese Aufforderung: Der Mensch soll in Verantwortung vor dem Schöpfer das von Gott Geschaffene verwalten und als Beauftragter Gottes so damit umgehen, dass das vom Menschen Entwickelte stets erkannt werden kann als etwas, das in Gott seinen Ursprung hat und auf Gott hin seine Ausrichtung hat. Anders und schlichter gesagt: Der Mensch soll sich der Welt gegenüber so verhalten, wie Gott sich gegenüber dem Menschen verhält. Oder wieder etwas differenzierter gesagt: Der Gedanke von der schöpferischen Erhaltung der Welt enthält die Aufforderung, „die Zukunft der Welt so zu gestalten, daß sie widerspruchlos ganz und gar Gottes Werk und ganz und gar unser Werk“<sup>8</sup> ist. Der Mensch handelt nicht an Gottes statt, sondern in Gottes Sinne. Wenn der Mensch also in der Welt als Geschöpf handelt, bedeutet dies, teilzunehmen am Entwicklungsprozess der von Gott geschaffenen Welt und diesen Entwicklungsprozess als freies und darin verantwortliches Subjekt zu gestalten.<sup>9</sup> Die theologische Rede von der Schöpfung hat also eine zentrale Pointe darin, dass sie gerade nicht irgendeinen einmal gegebenen Naturzustand einfrieren will, sondern im Gegenteil den Menschen auffordert, in Gottes Sinne weiterzuwirken. Die theologische Rede von der Schöpfung betont den Vorrang des Möglichen vor dem Wirklichen und fordert den Menschen dazu auf, sich an der Realisierung dieses Möglichen zu beteiligen – nicht an Gottes statt, aber in Gottes Sinne.

Bezogen auf unsere Fragestellung nach einer theologisch-ethischen Beurteilung von KI im Sozialwesen bedeutet dies zweierlei. Erstens, es ist per se legitim, dass Menschen sich an die Erfindung, Verfeinerung und immer effizientere Benutzung von KI machen. Das gehört grundsätzlich, im Sinne des christlichen Schöpfungsgedankens, zu den vorgesehenen Formen menschlichen Lebens in der Welt und menschlicher Gestaltung der Welt. Mit dem Schöpfungsgedanken ist die Begründung freien

---

8 Eberhard Jüngel: *Art. Schöpfung und Erhaltung 2. Dogmatisch*, in: RGG<sup>4</sup>, Bd. 7 (2004), Sp. 979f., 980.

9 Reiner Anselm: *Schöpfung als Deutung der Lebenswirklichkeit*, in: *Schöpfung (Themen der Theologie 4)*, hg. von Konrad Schmid, Tübingen 2012, S. 225–294, 267.

menschlichen Handelns in der Welt gesetzt. Schöpfungsgemäß handelt der Mensch, wenn er in Freiheit handelt. In Freiheit zu handeln und Freiheit zu erhalten, ist zugleich ein erstes Kriterium der Anwendung von KI. Ein zweiter Aspekt wird später hinzukommen, nämlich die Orientierung am Kriterium der Individualität. Beidem will ich mich jetzt nacheinander zuwenden.

### **1.2 Das Kriterium der Sicherstellung von Freiheit**

Der Aspekt der Freiheit des Menschen hinsichtlich der Schöpfung hat seine Pointe darin, dass Menschen gottgegebene Freiheit gegenüber dem Vorfindlichen empfinden.<sup>10</sup> Im Anschluss an Gen 1 lässt sich das schöpferische Handeln Gottes bestimmen als einen „die Welt auf Freiheit hin entwerfenden Akt ursprünglichen Anfangens“<sup>11</sup>. Er wird dadurch fortgesetzt, dass Menschen ein für die Welt „wohltätiges und heilsames“<sup>12</sup> Gegenüber Gottes werden, indem sie in Freiheit dem Verhältnis Gottes zu seiner Schöpfung entsprechen, so Röm 8. Als Geschöpf Gottes fortwährend und fortbauend in dessen Schöpfung zu handeln, schließt also ein, mit einer Zukunft zu rechnen, die nicht einfach nur eine „Extrapolation des Bestehenden darstellt“<sup>13</sup>. Als eine solche berechenbare Zukunft wäre sie gar keine echte Zukunft mehr. Es ist vielmehr eine Zukunft, die sich nicht auf das jetzt schon Wirkliche beschränkt, sondern auf das Mögliche weist. Und darum drängt sie den sich als Geschöpf Gottes verstehenden Menschen dazu, eine dem Möglichen entsprechende Gestaltung der Wirklichkeit anzustreben.

Das heißt zunächst: Innerhalb der von Gott geschaffenen Wirklichkeit ist ein Raum der Freiheit eröffnet, in dem sich Vertrauen zum Möglichen einstellen kann und soll.<sup>14</sup> Menschen, die sich als Geschöpfe Gottes verstehen und in Gottes Schöpfung schöpfungsgemäß handeln wollen,

---

10 Eberhard Jüngel: *Die Welt als Möglichkeit und Wirklichkeit. Zum ontologischen Ansatz der Rechtfertigungslehre*, in: Ders.: *Unterwegs zur Sache. Theologische Bemerkungen*, München 1988, S. 206–233, 226–231.

11 Jüngel: *Art. Schöpfung und Erhaltung* (s.o. Anm. 6), Sp. 979.

12 Michael Welker: *Art. Schöpfung*, in: *Wörterbuch des Christentums* (1988), S. 1119f, 1120.

13 Anselm: *Schöpfung* (s.o. Anm. 9), S. 258.

14 Jüngel: *Die Welt als Möglichkeit und Wirklichkeit* (s.o. Anm. 10), S. 229.

sollen und können sich in Freiheit diesen Möglichkeiten annähern. Das ist, nebenbei gesagt, übrigens auch ein Argument gegen die Ideologisierung bestimmter vorfindlicher Strukturen, wie er in ethischen Zuspitzungen der Schöpfungslehre immer wieder vorgenommen wird: Wer die Schöpfungslehre als Begründung für die Zementierung des Bestehenden heranziehen will, kann sich jedenfalls kaum auf Gen 1 oder auf Röm 8 berufen. Bezogen auf unsere Fragestellung nach KI im Sozialwesen heißt dies: Dass Menschen, die sich als Geschöpfe Gottes verstehen und in Gottes Schöpfung angemessen handeln wollen, die Möglichkeiten von KI nutzen und intensivieren wollen, ist zunächst vollkommen durch die im Schöpfungsgedanken eröffneten Freiheits- und Möglichkeitsräume legitimiert.

Allerdings ist das noch lange keine schrankenlose Freiheit. Eine zweite Pointe der Eröffnung von Freiheitsräumen liegt darin, dass sie auch die Unterscheidung von schöpfungsgemäßigem und eben nicht mehr schöpfungsgemäßigem Handeln anleitet. Die Eröffnung von Freiheitsräumen ist nicht nur eine Legitimation für die Nutzung und Intensivierung von KI, sondern zugleich auch ein limitierendes Kriterium für die schöpfungsgemäße Anwendung von KI. KI ist, um es zugespitzt zu sagen, nur dann schöpfungsgemäß eingesetzt, wenn sie selbst wiederum Freiheitsräume eröffnet oder zumindest offenhält. Bezogen auf unsere Frage nach KI im christlichen Gesundheits- und Sozialwesen mag das z.B. die Dokumentationssoftware sein, die die Spielräume zur Einstellung von Pflegekräften erhöht oder Möglichkeiten einer Ausweitung von deren genuinen pflegerischen Tätigkeiten eröffnet. Freiheitszugewinn mag z.B. auch das Kriterium für den Einsatz von Robotiksystemen sein. Umgekehrt aber wirkt das Kriterium des Zugewinns von Freiheit zum Beispiel dort limitierend, wo KI Kommunikation standardisiert. Ich komme auf all das später noch ausführlicher zurück. Jetzt aber zunächst zu dem zweiten oben genannten Kriterium.

### *1.3 Das Kriterium der Ermöglichung von Individualität*

Das zweite Kriterium, dem die menschliche Gestaltung der Welt im Sinne des Schöpfungsgedankens zu genügen hat, besteht in der Ermöglichung von Individualität. Individualität im Sinne der Schöpfungs idee ist nichts, das sich bereits aus dem Natürlichen ergäbe, weder als Feststehendes noch als etwas Vorprogrammiertes. Vielmehr hält der Schöpfungsgedanke fest, dass das Individuelle nicht von vornherein feststeht

und nicht festgelegt ist. „Gerade das Einzelne und Individuelle [ist] in seinem So-sein kontingent, nicht aus Naturgesetzen ableitbar“. Es erweist sich „aber genau in dieser Kontingenzt [...] als das von Gott Gewollte“<sup>15</sup>. Das betrifft zunächst einmal das Selbstverständnis eines jeden Einzelnen in der Welt: Er versteht sich in seiner Individualität als Resultat von Gottes Schöpfungshandeln. Es betrifft sodann aber auch das Verständnis, das der Mensch für seine Mitmenschen hat. In ihnen erblickt er ja ebenfalls Geschöpfe. Daraus ergibt sich die Norm des Umgangs mit den Mitmenschen. Auch sie sollen ihrer Geschöpflichkeit entsprechend leben können. Das heißt, es soll auch ihre Individualität ermöglicht und garantiert werden: ihre Individualität als das noch nicht Feststehende und niemals Abgeschlossene, stets im Werden Befindliche. Der Umgang von Menschen miteinander im Sinne des Schöpfungsgedankens bedeutet, dass man einander jenen individualitätsverfeinernden Wandel konzidiert und ermöglicht, der individuelle Entwicklung und individuelle Zukunft sicherstellt.

Damit haben wir das zentrale Kriterium, das für das Handeln des gottgeschaffenen Menschen in der gottgeschaffenen Welt und unter den Mitgeschöpfen gilt: Handle so, dass Individualität zugestanden und intensiviert wird. Das heißt: Handle so, dass Menschen und Dingen ihre individuelle Entwicklung und ihre individuelle Zukunft ermöglicht und dauerhaft offengehalten wird. Das gilt, um es zu konkretisieren und auf unsere Frage zu beziehen, dann fraglos auch für KI im christlichen Sozialwesen: Setze KI so ein und nur so ein, dass Menschen und Dingen ihre individuelle Entwicklung und ihre individuelle Zukunft ermöglicht und dauerhaft offengehalten wird. Umgekehrt: Wo der Einsatz von KI sich individualitätseinschränkend auswirkt, individuelle Entwicklung einengt, zum Beispiel durch Standardisierung oder Normierung, da wäre dieser Einsatz von KI fehl am Platze.

Ich bin fast am Ende dieses Abschnittes und fasse ihn zusammen: In der christlichen Schöpfungsvorstellung, jedenfalls in seiner evangelischen und neuzeitlichen Gestalt, ist angelegt, dass der als Geschöpf sich verstehende Mensch, der schöpfungsgemäß in der als Gottes Schöpfung verstandenen Welt handeln will und soll, sich in Freiheit den Möglichkeiten des Handelns zuwendet. Das gilt auch für die Nutzung und In-

---

15 Anselm: *Schöpfung* (s.o. Anm. 9), S. 258.

tensivierung von KI. Was der Mensch dann allerdings tut, in unserem Falle: wie er KI im Sozialwesen anwendet, muss sich an Kriterien messen lassen, die wiederum aus der Schöpfungsvorstellung ableitbar sind. Es handelt sich vor allem um zwei Kriterien, um dasjenige der Sicherstellung von Freiheit und um das der Ermöglichung von Individualität. Wo der Einsatz von KI diesen Kriterien entspricht, wird man sie als Verbesserung nutzen. Wo der Einsatz von KI jedoch jene Sicherstellung von Freiheit und Ermöglichung von Individualität zu gefährden droht oder gar konterkariert, da wird man sie im christlichen Sozialwesen auch nicht nutzen wollen.

## ***2. Konkretionen der kriteriengeleiteten Anwendung von KI im christlichen Gesundheitswesen***

Haben wir damit eine grundsätzliche Legitimation von KI im christlichen Sozialwesen und auch prinzipielle Kriterien für den Einsatz von KI in Sozialunternehmen, so bedarf dies doch der Konkretion: was ist unproblematisch, was ist problematisch, aber unter bestimmten Bedingungen bzw. in bestimmten Formen denkbar und was ist ausgeschlossen? Im folgenden Abschnitt möchte ich die genannten Kriterien auf verschiedene Anwendungsbereiche von KI beziehen und sie entsprechend sortieren.

### ***2.1 Unproblematische Anwendungsformen***

Ich beginne mit den aus meiner Sicht unproblematischen Formen der Arbeitsentlastungen, die KI in der Sozialarbeit ermöglicht. Dazu zähle ich erstens alle KI-gestützten *Automatisierungen von Verwaltungsroutinen* wie zum Beispiel die Unterstützung der Vorratsverwaltung, die Erstellung von Dienstplänen, aber auch die Erstellung von Verlaufsprognosen für den stationären Aufenthalt von Patienten und Patientinnen in der Klinik oder Pflegeeinrichtung. Sie dienen der Erleichterung genauso wie andere, ältere Hilfsmittel und es ist kaum zu sehen, wo sie Freiheit oder Individualität der Handelnden oder der Behandelten substantiell und kontraproduktiv einschränken könnten.

Das gleiche gilt, zweitens, meines Erachtens für KI-gestützte *Dokumentationssysteme* zur Dokumentation ärztlicher und pflegerischer Tätigkeiten. Sie werden vor allem in Kontexten der Abrechnung und

Haftung eingesetzt und zielen zunächst ja darauf, dass Pfleger und Pflegerinnen, Ärzte und Ärztinnen Zeit für ihre eigentlichen Tätigkeiten gewinnen – man könnte also sagen, dass sie sogar einen Freiheitszugewinn bedeuten. Möglicherweise erleichtert künftig die automatisierte Analyse von Patientendaten auch einmal die automatisierte Erstellung von individuellen Behandlungsplänen. Auch hier liegt das Problem weniger in einer Gefährdung von Freiheit und Individualität, allenfalls in dem allerdings technischen Problem der Preisgabe und Verwendung von Daten.

Damit sind Arbeitsentlastungen genannt, gegen die es aus theologisch-ethischer Sicht und unter Veranschlagung der genannten Kriterien der Erhaltung und Förderung von Freiheit und Individualität keine prinzipiellen Einwände gibt. Es handelt sich um funktionale Erleichterungen, die wir annehmen wie seinerzeit die Schreibmaschine, das Telefon oder den Personal Computer.

## ***2.2 Problematische, aber unter bestimmten Bedingungen legitime Anwendungsformen***

Komplexer wird es im Blick auf einen zweiten Typus des Einsatzes von KI im Sozial- und Gesundheitswesen, und zwar bei KI-Tools, die einerseits Arbeitserleichterungen im oben genannten Sinne darstellen, auf der anderen Seite aber dazu neigen, die Kommunikation zu entpersonalisieren und damit tendenziell zu standardisieren, man könnte auch sagen: zu entindividualisieren. Ich nenne zunächst einige solcher Tools, um dann die konkreten Herausforderungen im Einsatz dieser Tools und im Umgang mit ihnen zu beschreiben.

Welche Tools sind es, die in die Kommunikation eingreifen? Ich denke in erster Linie an *assistive Technologien* wie zum Beispiel an Sprachassistenten zur Entgegennahme von Wünschen der Patienten, an Interaktionssysteme, die Menschen mit eingeschränkter Alltagskompetenz bei der selbständigen Haushalts- und Lebensführung unterstützen – etwa durch die Erinnerung an notwendige Aktivitäten, an Medikamenteneinnahme – oder die die Information von Pflegenden leisten. Zu diesen assistiven Technologien zählen auch Formen der AI for Accessibility, die Menschen mit Behinderungen unterstützen, also etwa die Seeing AI, die Sehbehinderten Texte vorliest oder Gegenstände erkennt und beschreibt, die vor der Smartphonekamera erscheinen.

Zu den in die Kommunikation eingreifenden Tools könnte man neben den assistiven Technologien auch *Robotik-Systeme zur Unterstützung Pflegebedürftiger* beim Aufstehen, bei der Körperpflege oder bei der Nahrungsaufnahme zählen, außerdem auch *Sensorsysteme zur Überwachung von Gesundheitszuständen*. Mit gewissen Transferüberlegungen könnte man den Einsatz von *KI-basierten Lehr- und Lernsystemen in der Ausbildung* und Schulung von Pflegekräften auch in diese Rubrik rechnen.

Das gemeinsame Problem all dieser Anwendungen aus theologisch-ethischer Sicht und im Lichte der mich hier leitenden Kriterien besteht darin, dass sie auf der einen Seite passgenaue Hilfen für die Pflege von bedürftigen Menschen bereitstellen und dadurch die Verlängerung von Autonomie ermöglichen. Auf der anderen Seite aber treten sie, wo immer sie eingesetzt werden, stets auch an die Stelle menschlicher Zuwendung, menschlicher Fürsorge durch verbale und nonverbale Kommunikation. Letzteres ist noch kein Grund, diese Tools abzulehnen: In fast allen Fällen besteht die Alternative ja nicht zwischen KI-gestützten Tools und personaler Kommunikation, sondern zwischen KI-gestützten Tools oder gar keiner Unterstützung. Aber umgekehrt ist das noch lange kein Grund, diese Tools unkritisch anzuwenden.

Klarer wird nun aber zunächst das gemeinsame Problem dieser Anwendungen. Es besteht darin, dass es sich um maschinelle Kommunikation handelt, die zwar im Blick auf ihre funktionalen Wirkungen Freiheitszugewinne bedeuten kann und auch zur Verfeinerung individuellen Lebens dienen kann. Zugleich unterwirft sie aber die Kommunikation der unfreien Abhängigkeit von Algorithmen und entindividualisiert sie damit. Und damit sehen wir jetzt die Aufgabe beim Einsatz dieser Anwendungen klarer.

Diese Aufgabe besteht darin, im Einsatz dieser Anwendungen allen Beteiligten, so gut es geht, im Bewusstsein zu halten, dass es sich bei den kommunikativen Formen, in die die KI-Maschinen eintreten, um automatisierte, künstliche, nicht menschliche Kommunikation handelt, die die menschliche und personale Kommunikation nicht ersetzen kann oder ersetzen will – sondern die durch Inkaufnahme einer Simulation von Kommunikation freiheits- und individualitätsfördernde Erleichterungen bringen will. Anders und einfacher gesagt: Das Gerät ist kein echter Mensch, aber es erleichtert dir, ein echter Mensch zu sein. Das ist eine ziemlich komplexe gedankliche Operation, die man aber allen Beteiligten unter Veranschlagung ihrer individuellen reflexiven Belastbarkeit

so weit wie möglich zumuten muss, damit keiner der Beteiligten durch eine – und sei es schleichende, unbewusste – Verwechslung von maschineller mit personaler Kommunikation in die kommunikative Unfreiheit und Entindividualisierung getrieben wird.

Nun wird man an dieser heiklen und herausforderungsvollen Stelle allerdings gerade aus theologischer Sicht optimistisch sein können, dass es gelingt, die freiheits- und individualitätsfördernden Potentiale künstlicher, nichtpersonaler Kommunikation herauszustellen und zu nutzen. Ich will es an dem Kommunikationsroboter Navel<sup>16</sup> verdeutlichen. Wir sprechen mit ihm, als wäre er ein Mensch, der er natürlich nicht ist. Man muss zunächst einmal sagen: Das ist, als Ersatz, wünschenswerter als gar keine Kommunikation. Die maschinell simulierte Kommunikation ist im Falle der Kommunikation mit dementen Personen sogar besser, weil die Maschine geduldiger agiert als der menschliche Gesprächspartner. Man kann die Legitimität der Simulation auch damit begründen, dass man sie als Training für echte zwischenmenschliche Kommunikation beschreibt: Der Mensch trainiert in der maschinellen Kommunikation zwischenmenschliche Kommunikation, er bringt vielleicht in der künstlichen Kommunikation manche Dinge ins Bewusstsein und auch zum sprachlichen Ausdruck, die sonst nicht ins Bewusstsein und in die Artikulation gelangt wären.

Ersatz zwischenmenschlicher Kommunikation und Training zwischenmenschlicher Kommunikation sind die Stichworte, bei denen der Theologe ruhig wird. Denn tatsächlich haben wir damit in der religiösen Tradition ja Erfahrung und Erfolg. Ist doch jedes menschliche Gebet die Kommunikation mit einem Wesen, das, vorsichtig gesagt, von einer anderen Art von Realität ist als eine leibseelische menschliche Person. Aber wir kommunizieren mit diesem Wesen gleichwohl und bewusst und gewollt so, als wäre dieses Wesen eine menschliche Person. Wir tun das im Gebet ja nicht aus Gedankenlosigkeit oder aus Verlegenheit, dass wir mit Gott so reden, als wäre er eine Person. Wir tun das, weil wir in dieser Art simulierter menschlicher Kommunikation mit dem personal begriffenen Gott uns Dinge ins Bewusstsein bringen und Dinge in Worte fassen, die sonst unbewusst und ungesagt blieben. Wir wissen ja eigentlich, außerhalb des Vollzugs des Gebetes selbst, dass es ein unechtes

---

16 <https://navelrobotics.com/>

Gespräch ist, aber wir wollen dieses unechte Gespräch absichtlich, weil es uns lehrt und darin übt, zu hören und zu sprechen. Ob als Ersatz echter personaler Kommunikation oder als einübende Vorbereitung echter personaler Kommunikation ist dabei zweitrangig. Noch einmal anders gesagt: Im Gebet reden wir mit jemandem, von dem wir wissen, dass er nicht in der Weise wirklich ist, wie wir im Gebet uns selbst bewusst und willentlich vortäuschen. Nun: Warum sollte, was für Gott gilt, in *dieser Hinsicht* nicht auch für Navel gelten? Warum sollten wir nicht mit Navel reden oder reden lassen, als wäre er ein Mensch – wenn wir nur außerhalb dieser Gespräche mit Navel uns bewusst halten, dass er natürlich kein Mensch ist, sondern uns die Kommunikation mit einem Menschen ersetzt oder antrainiert?

Ich gehe von dem Ausflug zu Navel zurück zu dem in diesem Abschnitt diskutierten Typus von KI-Tools, die in die Kommunikation eingreifen. Die Herausforderung ist klar: Sie können Freiheit und Individualität fördern, tun dies aber um den Preis, dass sie unfreie und entindividualisierende, maschinelle Kommunikation betreiben. Und die Legitimität des Einsatzes dieser Tools im Sozialwesen ist von zwei Bedingungen abhängig: Erstens (es war oben schon gesagt), dass allen Beteiligten so eindringlich wie möglich, ihren jeweiligen Auffassungsgaben entsprechend, und immer wieder von Neuem diese Funktion der Tools vor Augen gehalten werden: dass sie simulierte Kommunikation betreiben, um freie und individuelle Autonomie zu schaffen oder zu erhalten. Und die zweite Bedingung besteht darin, dass diese Tools auch wirklich nur zu diesem Zweck eingesetzt werden: um freie und individuelle Autonomie zu schaffen oder zu erhalten. Niemals aber dürfen sie mit dem Zweck eingesetzt werden, authentische menschliche Kommunikation in all ihrer freien und individuellen Unberechenbarkeit zu ersetzen.

### *2.3 Ausgeschlossene, abzulehnende Anwendungsformen*

Ich komme drittens zu Formen des Einsatzes von KI im Sozial- und Gesundheitswesen, die meines Erachtens abzulehnen sind. Dies betrifft alle Formen des Einsatzes von KI-Systemen im Sozial- und vor allem im Gesundheitswesen, die die Beziehung zwischen Arzt und Patient dadurch tangieren, dass sie den Arzt in seiner Beratung abhängig machen von KI-erzeugten Daten und damit auch den Patienten in diese Abhängigkeit bringen. Das will ich erklären.

Zu denken ist in diesem Zusammenhang an das KI-gestützte *maschinelle Lernen anhand großer medizinischer Datensätze*, um Vorhersagen zu treffen und Entscheidungen zu indizieren. So lassen sich beispielsweise anhand großer radiologischer Bilddaten Krankheiten früher und zuverlässiger erkennen als mit noch so großer fachärztlicher Kenntnis. KI-gestützte Diagnosesysteme vermögen Symptome zu identifizieren, sie mit gespeicherten Daten zu vergleichen und auf dieser Grundlage Behandlungsempfehlungen auszusprechen. KI-Systeme können Krankheitsverläufe erkennen und Ärzten dabei zu helfen, Lücken ihrer Erkenntnisse aufzufüllen und sichere Diagnosen zu stellen. Das heißt: KI-gestützte Entscheidungssysteme nutzen Algorithmen des maschinellen Lernens, um Ärzte bei der Entscheidungsfindung zu unterstützen und personalisierte Behandlungsempfehlungen zu geben. KI-Systeme können gesundheitliche, genetische und auch soziale Daten einer Person so auswerten, dass die Fähigkeit eines Arztes, die optimale Therapie zu wählen, erheblich unterstützt werden kann.

So weit, so gut. Die Grenze scheint mir aber dort erreicht, wo KI nicht mehr nur unterstützt, sondern wo der Arzt in Abhängigkeit von KI-generierten Daten, Prognosen und Empfehlungen gerät. Denn das greift erheblich in die Beziehung zwischen Arzt und Patient ein.<sup>17</sup> Ist der Arzt abhängig, dann basiert seine Vertrauenswürdigkeit nicht mehr darauf, dass er sein Fachwissen mit Erfahrung und Augenmaß verbindet und darum aufgrund einer Gesamtinterpretation zu einem für den Patienten vertrauenswürdigen Therapievorschlagn kommt. Auch kann der Arzt keine eigene Verantwortung mehr für seinen Therapievorschlagn übernehmen, diese Verantwortung ist an die Technik delegiert. Zunächst der Arzt und dann mit ihm auch der Patient müssen den in ihren Gründen nicht mehr durchschauten Empfehlungen der KI vertrauen. Problematisch wird es also dort, wo es zu einer Umlenkung des notwendigen Vertrauens kommt. Es ist dann nicht mehr so, dass der Patient der Person des Arztes mit seinem Fachwissen, seiner Erfahrung, seinen ethischen Standards, seiner Fähigkeit zur Verantwortungswahrnehmung usw. vertraut. Vielmehr hat, unter der Dominanz von KI-Systemen, nicht mehr

---

17 Vgl. zum Folgenden: *Steering Committee for Human Rights in the fields of Biomedicine and Health (CDBIO): Report on the Application of Artificial Intelligence in Healthcare and its Impact on the 'Patient-Doctor'-Relationship (September 2024), Council of Europe, [https://health.ec.europa.eu/ehealth-digital-health-and-care/artificial-intelligence-healthcare\\_en](https://health.ec.europa.eu/ehealth-digital-health-and-care/artificial-intelligence-healthcare_en)*

der Patient das Vertrauen in den Arzt, sondern der Arzt und mit ihm der Patient müssen gemeinsam Vertrauen in die technische Leistung und Zuverlässigkeit der KI-Systeme haben.

Diese notwendige Umlenkung des Vertrauens ist meines Erachtens problematisch dort, wo sie zum blinden Vertrauen werden muss. Dies ist anders angelegt als in anderen und üblichen Formen blinden Vertrauens in die Zuverlässigkeit der Technik. Denn hier ist die Behandlungsempfehlung des Arztes keine selbst verantwortete mehr, sondern eine automatisierte Entscheidung. Die Behandlungsempfehlung des Arztes ist damit aus dem Bereich menschlicher Beziehungen herausgenommen. Menschliche Beziehungen aber konstituieren einen Raum der Freiheit, der individuell gestaltet werden muss, zugleich aber permanent bedroht ist und darum verantwortlich gestaltet werden muss. Das notwendig blinde Vertrauen in die Empfehlung von KI-Systemen aber verbietet die von ihm selbst verantwortbare Entscheidung des Arztes oder später dann die Entscheidung des Patienten aufgrund transparenter und vertrauenerweckender Gründe. Vielmehr erfordert es die Haltung der Unterwerfung. Das aber ist, als Haltung des Arztes und später dann des Patienten, ersichtlich das Gegenteil der in den Schöpfungsvorstellungen angelegten Idee der individualitätsfördernden Ermöglichung von Freiheit. Insofern sind, aus der Sicht christlicher Ethik, meines Erachtens die genannten Entscheidungshilfen der KI-Systeme problematisch in dem Augenblick, in dem der Arzt sie nicht mehr unterstützend nutzt, sondern in dem seine Fähigkeit zur autonomen Dateninterpretation und anschließend die Fähigkeit des Patienten zur autonomen Nutzung dieser Dateninterpretation eingeschränkt ist. Der Preis ist meines Erachtens zu hoch: Arzt und Patient werden unwiderruflich abhängig von Intransparenz.

### 3. *Schluss*

Ich komme zu einem ganz kurzen Schluss meiner Überlegungen. Deutlich ist und wenig überraschend: Es ist auf der einen Seite nicht nur unvernünftig, sondern widerspräche auch dem Effizienzideal in der christlichen Sozialarbeit, das Unterstützungspotential von KI zu ignorieren. Auf der anderen Seite gibt es Grenzen der Eingriffstiefe, die man KI zubilligen möchte. Darum bringt KI für das christliche Sozial- und Gesundheitswesen auf der einen Seite erhebliche Erleichterungen mit sich, erhöht zugleich auf der anderen Seite aber den Bedarf nach kriteri-

engeleiteter Reflexion der ethisch vertretbaren Anwendung. Das ist eine unabschließbare Aufgabe angesichts der zu erwartenden Fortschritte im Bereich der Entwicklung von KI. Sie ist stets von neuem in Angriff zu nehmen und muss sich mit immer neuen Möglichkeiten sowie Grenzen befassen. Auch stehen die Kriterien des Einsatzes von KI nicht ein für alle Mal fest, sondern müssen stets von neuem überprüft und verfeinert werden. Die ethische Reflexion, der Aufbau eines sicheren ethischen Urteils ist eine, so scheint mir, an Bedeutung und an Umfang zunehmende Aufgabe für Führungskräfte diakonischer Einrichtungen und Unternehmen. Gerade diese brauchen für die Ermittlung des vertretbaren Einsatzes von Künstlicher Intelligenz ein hohes Maß an natürlicher Intelligenz.

# IMPRESSUM

Christian Albrecht [Hrsg.]:  
**Künstliche Intelligenz in diakonischen Unternehmen**

Diakonie reflektiert – Band 3 (2025)

Eine Reihe der Evangelischen Arbeitsstelle für missionarische  
Kirchenentwicklung und diakonische Profilbildung (midi)

*Alle Rechte vorbehalten – All rights reserved*

**[www.diakonie-reflektiert.de](http://www.diakonie-reflektiert.de)**

**[www.mi-di.de](http://www.mi-di.de)**

Augustinum  $\Phi$

