Ralf Stapelfeldt

Einleitung Teil 2: Grenzen und Folgen Künstlicher Intelligenz

Computer und Programme, denen wir im allgemeinen Sprachgebrauch eine Künstliche Intelligenz zuschreiben, sind seit den theoretischen Überlegungen in den 50er und 60er Jahren des letzten Jahrhunderts über die Jahrzehnte nicht nur realisiert worden, sondern gerade seit Beginn des neuen Jahrtausends auch immer mehr in der unmittelbaren Lebenswelt der Menschen angekommen. In Suchalgorithmen im Internet, Robotersystemen in der Arbeitswelt, Computerspielen oder Mustererkennungsprogrammen – um nur wenige Beispiele zu nennen - hat das, was wir ‚Künstliche Intelligenz' nennen, Einzug gehalten in unser tägliches Leben. KI-basierte Lösungen sind in der Arbeitswelt ‚en vogue' und versprechen z.B. in der automatisierten Kundenkommunikation, in der medizinischen Diagnostik oder in autonomen Fahrzeugen ein im Vergleich zum Menschen bei weitem überschreitendes Maß an Genauigkeit und Effizienz. In der laufenden Berichterstattung ist die Künstliche Intelligenz längst kein Randthema mehr für Nerds, sondern ins Zentrum aktueller und breit geführter Debatten gerückt. Künstliche Intelligenz weist dabei auf eine verheißungsvolle wie faszinierende und zugleich auch beängstigende Zukunft hin, in der sie unser Leben für immer verändern wird. Utopische und dystopische Visionen dazu bilden die Basis ungezählter Science-Fiction-Filme und Serien, die auf eine große Fangemeinde zählen können.

Was aber ist es, was eine Maschine, einen Computer oder einen Menschen überhaupt intelligent macht? Diese Frage, und das hat bereits der erste Teil dieses Bandes gezeigt, ist weitaus schwerer zu beantworten, als man zunächst erwarten würde. Man kann hier etwa auf die Fähigkeit des Denkens oder auf eine bestimmte Art des Handelns abstellen. Als Maßstab für das, was als intelligent zu betrachten ist, kann beispielsweise die Vernunft oder aber der Mensch selbst herangezogen werden (vgl. Russell / Norvig 2016, S. 2). Der Duden verbindet diese Gedanken und definiert Intelligenz als die „Fähigkeit [des Menschen], abstrakt und vernünftig zu denken und daraus zweckvolles Handeln abzuleiten" (Duden online 2020). Schon in seinem legendären Aufsatz von 1950 „*Computing machinery and intelligence*" schlägt Alan Turing vor,

das menschliche Denken und das dadurch mögliche Handeln - bei ihm am Beispiel sprachlicher Kommunikation - als Maßstab zu wählen. Um zu klären, ob Maschinen wie Menschen denken können, schlägt er sein berühmtes Imitationsspiel vor: Wenn eine Maschine in einer schriftlichen Unterhaltung so gute Antworten gibt, dass ein Dritter nicht mehr erkennen kann, ob er mit einer Maschine oder einem Menschen spricht, dann müssten wir ihr zugestehen, dass sie wie ein Mensch denken kann und ihr demnach Intelligenz zuschreiben (vgl. Turing 1950). Dieser Ansatz, dass sich Künstliche Intelligenz daran festmachen ließe, ob eine Maschine den gleichen Output wie ein Mensch nachahmen kann, ist seit den Anfangstagen der Fachdisziplin der Künstlichen Intelligenz bis heute stark verbreitet.[1]

Doch ebenso lange wird auch darüber diskutiert, ob Turing mit der Substitution seiner Frage nach der Denkfähigkeit von Maschinen durch sein Imitationsspiel nicht gerade den Wesensbereich menschlicher Intelligenz verlässt, die ursprüngliche Frage also nicht bloß reformuliert, sondern schlicht durch eine ganz andere ersetzt (siehe hierzu etwa den Aufsatz von Zimmerli in diesem Band). Seit den 50er Jahren wird darüber debattiert, ob die Simulation eines intelligenten Outputs ausreichend sein kann für die Frage, ob ein System intelligent ist oder nicht und ob zur menschlichen Intelligenz nicht gerade etwas hinzuzuzählen ist, das über einen entsprechenden Abgleich von Outputs gar nicht zu erfassen ist. Mit dieser Kritik an der Verifikationsmethode ist eine Kritik an der Idee, dass Maschinen prinzipiell das Niveau menschlicher Intelligenz überhaupt erreichen könnten, eng verbunden. Die Idee, dass dies prinzipiell erreichbar *ist*, wird einem Vorschlag von John Searle folgend auch als starke KI-These bezeichnet, während die Kritiker nur einer schwachen KI-These zustimmen können, wonach artifizielle Systeme lediglich Teilaspekte menschlicher Intelligenz erfolgreich simulieren können, aber eben nie im vollen Umfang als intelligent zu bezeichnen wären. Searle gehört zu den Kritikern der starken KI-These und meint, dass es Maschinen und Computer, selbst wenn es ihnen gelänge, uns etwa im Rahmen des Turing-Tests in der Kommunikation zu täuschen, es ihnen eben doch an etwas Entscheidendem fehle: In der Maschine sei niemand, der versteht, was sie sagt, niemand, der

1 Als Startschuss der Fachdisziplin der Künstlichen Intelligenz wird zumeist die Dartmouth Konferenz von 1956 genannt (vgl. die Beiträge von Christian Freksa, Hans-Jörg Kreowski und Wolfgang Krieger, Nadine Schumann und Yablo Du, sowie Walther Zimmerli in diesem Band). Die Basisannahme dieser Tagung lautete, dass jeder Aspekt menschlicher Intelligenz so exakt beschrieben werden könne, dass prinzipiell auch eine Maschine ihn *simulieren* kann (vgl. Russell / Norvig 2016, S. 17).

dem Gesagten eine Bedeutung zukommen lassen könne. Es fehlen ihr mithin Selbstbewusstsein und Intentionalität. Erst der Umstand, dass wir Wesen sind, die sich ihrer selbst bewusst sind, mache uns zu denkenden Wesen (vgl. Searle 1980, S. 417 ff.; Searle entwickelt in diesem Zusammenhang auch sein legendäres Chinese-Room-Argument, vgl. dazu die Einleitung zum 1. Teil und die Beiträge von Dieter Mersch und Jan Tobias Fuhrmann in diesem Band). Eine Maschine könne demnach niemals denken, weil schlicht kein Denker vorhanden ist.

Diese Art von Kritik, die letztlich auf einen unüberbrückbaren Wesensunterschied zwischen Geist und Materie verweist, ist eng verwandt mit der Auffassung, dass im menschlichen Geist etwas vor sich ginge, das nicht programmierbar sei, sich nicht in Algorithmen abbilden ließe und mithin nicht künstlich erzeugt werden könne. Der Philosoph J.R. Lucas argumentiert schon 1961 in einem einflussreichen Aufsatz, dass die weithin geteilten Einsichten des Mathematikers Kurt Gödel zum Schluss führten, dass der menschliche Geist prinzipiell die Möglichkeiten eines deterministisch-mechanistischen Systems, wie ein Computer, übersteige. Gödels berühmter Unvollständigkeitssatz von 1931 besagt, dass sich zu jeder formalen, axiomatischen Theorie unentscheidbare Sätze konstruieren lassen, die genau dann wahr sind, wenn sie zugleich innerhalb der Theorie unbeweisbar sind (vgl. Gödel 1931, siehe dazu auch die Beiträge von Dieter Mersch und Daniel Wenz in diesem Band). Diese unumstrittene mathematische Erkenntnis wendet Lucas auf den menschlichen Geist an und kommt zu dem Schluss, dass Gödels Theorem ein Beweis dafür wäre, dass der Geist nicht als eine Maschine verstanden werden könne. Wenn zu jeder formalen Theorie Gödel-Sätze gebildet werden können, die innerhalb dieser Theorie unbeweisbar sind, so folge daraus, dass eine Maschine, also etwa ein Computer oder ein anderes artifizielles System, einen solchen Satz prinzipiell nicht beweisen, und daher im Gegensatz zum intelligenten Menschen nicht als wahr erkennen könne (vgl. Lucas 1961, S. 112 f.).

Demnach wäre es egal, welchen Output eine Maschine auch produzieren mag, sie würde niemals den Grad menschlicher Intelligenz erreichen. Selbst dann, wenn sich eine Maschine zu philosophischen Fertigkeiten emporschwänge, wenn sie, um mit Aristoteles zu sprechen, „das lebendige Wirken des philosophischen Geistes“ (Aristoteles 1969, Buch X, 1177a24) zeigte und somit die höchste Trefflichkeit, die oberste Kraft des Menschen, nämlich die Tätigkeit der Vernunft (vgl. ebd., 1177a13) simulieren könnte, bliebe ihr gleichwohl das Prädikat ‚*wirklich* intelligent‘ verwehrt. Man könnte es auch folgendermaßen ausdrücken: „Philosophen haben jahrzehntelang versucht, Computer wie

Menschen denken zu lassen. Das Problem ist, dass sich das nicht gut machen lässt, weil menschliches Denken Konzepte berührt wie Intentionalität, Agentivität und so weiter. Wenn man versucht einen Computer mit solchen Dingen zu programmieren, wird das niemals richtig funktionieren." (zitiert nach Tremmel 2020). Das Kuriose an diesem Zitat ist, dass die darin aus Sicht vieler Kritiker kluge Einsicht wiederum, in erstaunlicher Wendung gegen sich selbst, von einem KI-Sprachmodul namens GPT-3 (*Generative Pre-Trained Transformer 3*, siehe auch den Beitrag von Elektra Wagenrad in diesem Band) generiert wurde, das mit dem Inhalt eines philosophischen Blogs gefüttert wurde und dann selbst an der Debatte teilnahm. Von GPT-3 stammt auch die Aussage, dass „das Embodiment eines Systems in der Welt" (ebd.) ein wichtiger Aspekt der Intelligenz sei, was direkt zu einer anderen einflussreichen Kritik an der Vorstellung führt, dass eine künstliche die menschliche Intelligenz vollumfänglich abbilden könnte.

Sie wird etwa von Hubert Dreyfus Anfang der 1970er Jahre geäußert. Zunächst weist dieser darauf hin, dass die These vollumfänglicher Künstlicher Intelligenz (in der heutigen internationalen Diskussion meist als *Artificial General Intelligence* bezeichnet) an starken Voraussetzungen geknüpft sei. Sie unterstelle, dass das menschliche Denken auf einen informationsverarbeitenden Prozess zu reduzieren sei, der digital nach formalen Regeln ablaufe und somit intelligentes menschliches Verhalten formal beschrieben werden könne, um es durch ein künstliches System zu erzeugen. Gerade das hält Dreyfus aber für empirisch höchst unwahrscheinlich und a priori für inkonsistent und selbstwidersprüchlich. Vielmehr müssten wir davon ausgehen, dass das menschliche Verhalten nicht programmierbar und die menschliche Intelligenz nicht formalisierbar sei. (vgl. Dreyfus 1972, S. 197 f.) Für Dreyfus kann Künstliche Intelligenz nicht die Einbettung des Menschen in einer Welt erfassen, denn es mangele ihr an einem lebenden Körper und einer menschlichen Sozialisation. Die Verkörperung und das In-der-Welt-sein, wie es Heidegger nennt (vgl. Heidegger 2006, S. 62), werden zu konstitutiven Elementen menschlicher Intelligenz, die nicht künstlich reproduziert werden können (vgl. Misselhorn 2018, S. 27). Eng verbunden mit dieser Kritik ist die sogenannte *Embodiement* These[2], wonach Bewusstsein oder das Denken prinzipiell an eine Verkörperung gebunden sind, was übertragen auf die KI-Debatte hieße, dass auch Intelligenz einen lebenden Körper braucht, um zu entstehen.

2 Einen guten Einstieg in die Embodiement-Diskussion bietet z.B. der Sammelband „Philosophie der Verkörperung" (Fingerhut et al. 2013).

Die Kritik an der starken KI-These ist in der philosophischen Debatte nicht ohne Widerspruch geblieben. Ein prominenter Vertreter des Gegenlagers ist Daniel Dennett (vgl. auch seinen Beitrag in diesem Band). Er hält eine starke Künstliche Intelligenz zwar nicht für wünschenswert (vgl. Dennett 2019), aber für prinzipiell möglich (vgl. Dennett 1998). Das selbstbewusste ‚Ich' etwa wird von ihm als ein Trick des Gehirns gedeutet, als eine Abstraktion, die aus der Biografie des Körpers zusammengestellt wird, dessen narrativer Schwerpunkt sie sei (vgl. Dennett 1993, S. 426 f.). Es gibt für ihn kein ‚Ich', das im Inneren des Kopfes einem Bewusstseinsstrom folgte und in einem Computer deshalb immer fehlen würde: Auch beim Menschen entsteht Bewusstsein über einen informationsverarbeitenden Prozess im Gehirn, weshalb der Bewusstseinsstrom als Ergebnis dieses Prozesses nicht einem bereits vorhandenen Selbst wie in einem Filmtheater (bei Dennett das ‚Cartesianische Theater') vorgeführt werden könne. Es sei schlicht niemand da, der zuschaut. (vgl. Dennett 1993, S. 107 und S. 127). Intentionalität ist für Dennett eine externe Zuschreibung von Zuständen, die er auch einem Schachcomputer zubilligt (vgl. Dennett 1978, S. 4 ff.). Computer und Künstliche Intelligenzen können nach Dennett zwar nicht aus ihrem Programm herausspringen, weshalb sie auch, dem Unvollständigkeitssatz Gödels gemäß, im streng formalen Sinne nicht fähig sind, entsprechend formulierte Sätze zu beweisen. Diese Begrenzung gelte jedoch in gleicher Weise für den Menschen. Es sei etwas anderes, einen Satz als wahr zu erkennen, als ihn zu beweisen (vgl. Dennett 1995, S. 431).

Kritiker der starken KI-These führen Gefühle, Intuition oder die Gebundenheit an eine Verkörperung als Aspekte an, die allein der menschlichen Intelligenz vorbehalten seien. Die Apologeten der starken These halten dem entgegen, dass schon heute eine verblüffend realistische Expression von Gefühlen künstlicher Gesichter auch bei Robotern Einzug gehalten hat oder dass diese Roboter der menschlichen Verkörperung teilweise zumindest im Aussehen und in den Bewegungen bereits erstaunlich ähneln. Eine andere faszinierende Facette der schnellen Entwicklung im Umfeld artifizieller Intelligenz ist das derweil auch bei KI-Systemen aufkommende Polanyi-Paradox, das eigentlich in Bezug auf Menschen besagt, dass unser Wissen und unsere Fähigkeiten zu einem großen Teil jenseits expliziten Verstehens liegen. Dank neuronaler Netze und *Deep Learning* gelangen nun aber auch Computer zu einem Wissen, das niemand – weder das System noch die Programmierer des Systems - explizieren kann (vergleiche dazu auch den Beitrag von Daniel Wenz in diesem Band).

Gleichwohl sind die zwei aufgeführten Kritikpunkte an der Vorstellung einer starken KI bis heute extrem wirkmächtig:

1.: Künstliche Intelligenz könne niemals ‚wirklich' intelligent sein, weil das selbstbewusste ‚Ich' fehle, an dem nicht nur das Denken i.e.S., sondern wahlweise auch Intentionalität, Emotionen, phänomenale Bewusstseinszustände, freier Wille, Intuition und Ähnliches mehr hingen.

2.: Künstlicher Intelligenz fehle es an einer Verkörperung und menschlichen Sozialisation, sie sei nicht eingebettet in eine Welt, in der sie lernen und mit der sie reziprok und kontinuierlich in Verbindung steht, was aber konstitutiv für die menschliche Art von Intelligenz sei.

Diese Art der Kritik und der Gegenentwurf von Befürwortern der starken KI-These wie Dennett zeigen, dass die philosophische Debatte um Künstliche Intelligenz untrennbar mit den ganz zentralen Fragen der Philosophie des Geistes der letzten Jahrhunderte verbunden ist. Die KI-Debatte wirft Fragen auf, deren Antworten von den Grundpositionen abhängen, die in ihr eingenommen werden. Reese folgend sind es drei große Fragen, die das Feld kartographieren (vgl. Reese 2018, S. 39 ff.) und einen Einblick in die möglichen Standpunkte geben, die sich schon in den Aufsätzen des 1. Teils und ebenso so auch wieder in jenen des folgenden 2. Teils dieses Sammelbandes erkennen lassen. Die erste Frage lautet, was das Selbst eigentlich ist, das sich in unserem Geist seiner selbst bewusst wird: Ist es eine Seele, also eine vom Körper zu trennende Substanz, oder handelt es sich um ein stark emergentes Phänomen, das jenseits physikalischer Gesetzmäßigkeiten aus dem Zusammenspiel unseres Gehirns entsteht oder aber ist es schlicht ein besonders kluger Trick der Evolution, der im Rahmen darwinistischer Prozesse entstanden und in keiner Weise mysteriös ist, sondern auch physikalisch erklärt werden kann. Die zweite ist die metaphysische Frage nach unserer Realität, ob das Sein also monistisch oder dualistisch zu verstehen ist, ob es letztlich aus den physikalischen Elementarteilchen und den zwischen ihnen wirkenden Kräften besteht, oder ob im Falle des menschlichen Geistes etwas Spirituelles hinzukommt. Und die dritte schließlich ist die anthropologische Frage nach dem, was der Mensch ist: Ist er etwas Herausgestelltes und Besonderes, das etwa kraft seiner Vernunft allem anderen Leben enthoben ist, oder ist er ein evolutionär besonders weit entwickeltes Tier oder aber ist er letztlich auch nur eine Maschine? Je nach Position zu diesen drei Fragen ergibt sich ein anderer Raum von Optionen, wie die starke KI-These zu beurteilen ist.

Im zweiten Kapitel dieses Bandes finden sich interessante Aufsätze, die auf unterschiedliche Weise Position zur starken KI-These beziehen,

verschiedene Aspekte dazu beleuchten und somit einen guten Blick auf die aktuelle Debatte eröffnen. Doch in diesem zweiten Teil steht neben der Frage, was Künstliche Intelligenz eigentlich ist und wie sie metaphysisch und anthropologisch gedeutet werden kann, auch die Frage im Mittelpunkt, welche Folgen sich daraus für den Menschen und die Gesellschaft, in der er lebt, ergeben. Denn die rasante Entwicklung von KI-Systemen führt nicht nur zur philosophischen Herausforderung, entweder die menschliche von der Künstlichen Intelligenz abzugrenzen (für Vertreter der schwachen KI-These) oder aber auch den Geist und seine intelligenten Fähigkeiten letztlich als Ergebnis eines symbolverarbeitenden Prozesses zu verstehen (wozu sich Vertreter einer starken KI-These bekennen müssen), sondern auch zu praktischen Konsequenzen und ethischen Fragestellungen, wenn der Mensch immer häufiger mit KI-Systemen interagiert, die immer stärker an die Fähigkeiten des Menschen heranreichen.

Wenn Künstliche Intelligenz etwa eingesetzt wird, um auf Basis historischer Daten im Abgleich zu einem konkreten Einzelfall medizinische Diagnosen zu generieren, dann führt dies zu der weitreichenden Fragestellung, wie diese ‚Expertenmeinung' eines KI-Systems einzuordnen, wie mit ihr umzugehen ist, und wie sich Verantwortung in einem Prozess verteilt, in dem KI eine zentrale Funktion einnimmt. Ein anderes Beispiel sind KI-gesteuerte Systeme, die schon heute ein erstaunlich hohes Maß an Aktionsautonomie und eine hohe Vielfalt an Reaktionsmöglichkeiten erreicht haben. Autonom fahrende Autos oder militärische Drohnen sind bereits in der Realität angekommen und stellen uns weitreichende ethische Fragen. Wer trägt etwa die Verantwortung für die Aktionen (um nicht zu sagen ‚Handlungen') solcher autonomen Systeme, wenn sie im Zuge des oben schon erwähnten Polanyi-Paradox von niemandem mehr in Gänze verstanden werden. Die alte Weisheit, dass Computer immer nur exakt das tun, was ein Programmierer zuvor explizit eingegeben hat, gilt in dieser einfachen Form bei selbstlernenden Systemen nicht mehr. Und dort, wo eine exakte Programmierung vorgenommen wird, stellt sich die ethische Frage des Inhalts dieser Programmierung, also wie sich etwa ein autonom fahrendes Auto in Grenzsituationen, in denen ein Unfall nicht mehr zu vermeiden ist, verhalten soll.

Ein ganz anderes Feld, das sich auftut, wenn die Interaktion zwischen Mensch und KI in den Fokus genommen wird, sind die Folgen für das soziale Miteinander und Normen des Zusammenlebens. Was folgt beispielsweise aus dem Umstand, dass KI-Systeme zu selbstverständlichen Gesprächspartnern werden (etwa Apples Siri oder Ama-

zons Alexa) oder als Haushalts- oder Pflegeroboter Menschen nicht nur gute Dienste erweisen, sondern mit ihnen auch im Rahmen sozialer Bezugssysteme interagieren und zu ihnen eine Art von Vertrauensverhältnis aufgebaut wird? Handelt es sich bei diesen Dienstrobotern noch um reine Maschinen, die insofern auch als solche zu behandeln sind, oder überschreiten sie bereits die Schwelle zu einem moralischen Objekt, dem wir im Rahmen von Verhaltensregeln und Normen des sozialen Miteinanders begegnen? Vertreter der starken KI-These halten das Erwachen von Bewusstsein in solchen künstlichen Systemen für denkbar, in dessen Folge auch Aspekte wie Leidensfähigkeit, phänomenale Wahrnehmungen oder die Entwicklung eines Selbstbewusstseins mit all den sich daran anschließenden Konsequenzen Teil der zu führenden Diskussion werden. Doch es braucht keine so weitreichende Annahme, um zu erkennen, dass eine Diskussion um die Frage des richtigen Verhaltens gegenüber solchen sozialen Robotern (*social bots*) zu führen ist. Es könnte sein, dass ein gewohntes Verhalten gegenüber einem Roboter auch auf reale Lebewesen übertragen wird. Die Unterscheidung der subkutanen sozialen Normen wird umso verschwommener, je mehr die Interaktion zwischen Mensch und Roboter dem Miteinander unter Menschen ähnelt. Je mehr die *social bots* also das Aussehen, das Verhalten, die Sprechweise oder Gesten und Mimik eines echten Menschen nachahmen, desto eher erfolgt eine Anthropomorphisierung, die die sozialen Verhaltensnormen verschwimmen lässt und jene gegenüber Menschen auf Roboter überträgt oder eben umgekehrt. Die Gefahr, dass schädliches Verhalten aus Aktionen mit sozialen Bots auf reale Personen übertragen werden könnten, ist zumindest zu diskutieren.

Dieser Sammelband bietet dem/der interessierten Leser*in die Möglichkeit, in die oben angerissenen philosophischen Fragestellungen einzutauchen und einen spannenden Teil der sich daraus ergebenden aktuellen Debatte zu verfolgen, für die die facettenreichen, abwechslungsreichen und lesenswerten Aufsätze des zweiten Teils stehen.

Hans-Jörg Kreowski und **Wolfgang Krieger** fragen in ihrem Aufsatz „Künstliche Intelligenz – ‚künstlich' Ja, ‚Intelligenz' wohl kaum" zunächst, ob sich Intelligenz allein am Output oder auch an den Wirkprinzipien festmachen lasse. Nach ihnen kann etwas, das nach außen „wie echt" wirkt, gleichwohl etwas vollkommen Anderes sein. So, wie eine künstliche Blume zwar aussieht wie eine Blume, ihr aber zentrale Eigenschaften wie lebendig, natürlich gewachsen oder Fortpflanzungsfähigkeit fehlen, so sei „Künstliche Intelligenz" eben etwas ganz anderes als die Menschliche und nicht *wirklich* intelligent. Kreowski und Krieger weisen zudem auf konkrete Risiken hin, die sich aus dem

„Hype um KI“ ergeben und die sie mit dem Verweis auf den möglichen Verlust von Arbeitsplätzen, eine zunehmende soziale Überwachung und perfider werdende Waffensysteme konkretisieren.

Elektra Wagenrad stellt in ähnlicher Weise in ihrem Text „Das clevere Pferd Hans und die Blackbox der KI“ heraus, dass KI weder lebendig sei, noch Emotionen besitze oder einen Eigenwillen habe. Ähnlich wie ein Pferd, das im Zirkus vorgibt, rechnen zu können, wirke es bei der Künstlichen Intelligenz lediglich nach außen hin so, als würde sie die Probleme, die sie löst, verstehen, ohne dass dies tatsächlich der Fall sei.

Auch im Text „Vorbemerkungen zu einer Kritik algorithmischer Rationalität. Denken, Kreativität und Künstliche Intelligenz“ von **Dieter Mersch** kommt dieser zum Schluss, dass es einen Teil des menschlichen Denkens gibt, der nicht algorithmisch rationalisiert werden könne und sich damit einer Erfassung durch Künstliche Intelligenz entziehe. Das Problem rühre daher, dass im Zuge der Entwicklung der Künstlichen Intelligenz in gewisser Weise vergessen wurde, worauf alle informationstechnischen Prozeduren letztlich beruhen, nämlich auf mathematischen Modellierungen. Diese Modelle implizierten jedoch, so Mersch, Einschränkungen, z.B. der Diskretisierung von Variablenwerten, denen das menschliche Denken zumindest nach bisherigem Wissen nicht unterliege. Dies wirke sich besonders deutlich in der Unfähigkeit Künstlicher Intelligenz zu kreativen Prozessen aus.

Einen anderen Aspekt menschlicher Intelligenz, der durch eine KI nicht erfasst werden könne, ist nach **Gergana Vladova** und **Sascha Friesike** die Fähigkeit, sich zu irren und daraus zu lernen. In ihrem Aufsatz „Irren bleibt menschlich: Wieso falsche Entscheidungen für uns so wichtig sind und welche Rolle Künstliche Intelligenz dabei nicht spielen kann“ stellen sie die besondere Bedeutung des Lernens aus Fehlern für die menschliche Intelligenz heraus, während es bei der Künstlichen Intelligenz gerade darum gehe, jeden Fehler zu vermeiden.

Auch **Sybille Krämer** verweist in ihrem Aufsatz „Nüchtern bleiben! Künstliche Intelligenz jenseits des Mythos“ auf die Unterscheidung zwischen starker und schwacher Intelligenz hin, wobei sie diese als Bipolarität zwischen visionärer, geradezu mythischer KI einer fiktionalen Zukunft und einer zeitgenössisch-alltäglichen, effizienten und prosaischen KI erkennt und den Fokus auf letztere lenkt. Wenn die menschliche Kultur über Schrift, Bild, Text oder Diagramm seit jeher die Technik der Verflachung vollziehe und so die dreidimensionale Welt auf eine zweidimensionale Repräsentation reduziere, führe die Digitalisierung nun zu einer noch stärkeren Reduktion auf das allein Maschinenlesbare, wodurch ein computergeneriertes, digitales Schattenbild der Welt

entstehe. Krämer beleuchtet in ihrem Aufsatz daran anschließend Probleme der heutigen KI und insbesondere des Maschinenlernens. Schließlich beantwortet sie die schon von Turing aufgeworfene Frage, ob Maschinen intelligent sein können, indem sie sie umformuliert: „Können Maschinen Bedeutungszusammenhänge verstehen?" Ihre Antwort verweist zurück auf die bipolare Unterscheidung: Der Mensch versteht Sinnzusammenhänge, wo heutige KI ein lediglich oberflächliches, operatives Verstehen zeigt. Für Krämer deutet nichts darauf hin, dass Maschinen jemals das menschliche Verstehen von Sinn und Bedeutung werden leisten können.

Rico Hauswald meint in seinem Aufsatz „Digitale Orakel? Wie Künstliche Intelligenz unser System epistemischer Arbeitsteilung verändert.", dass durch neuronale Netze und *Deep Learning* die KI-Systeme immer mehr zu einer Black Box würden, wobei sie umso unverständlicher für uns Menschen würden, je smarter sie sind. Durch die immer weiter steigenden Fähigkeiten der KI sei sie zwar eine wichtige Unterstützung für Experten und im Zuge der Demokratisierung von Wissen praktisch für Laien, jedoch bestehe das Risiko paternalistischer Tendenzen, wenn der Mensch zunehmend seine eigenen Fertigkeiten im Zuge der Abgabe von Aufgaben an KI-Systeme und seine eigene Kritikfähigkeit gegenüber diesen aufgrund fehlenden Verständnisses verliere.

In eine ganz andere Richtung führt uns **Thomas Weiß**. Er diskutiert in „Künstliche Intelligenz – ein marxistisch-ökonomischer Blick" vor dem Hintergrund einer marxistischen Wirtschaftstheorie die Folgen eines fortschreitenden Einsatzes von KI in der Arbeitswelt bis hin zum logisch möglichen Endpunkt einer Vollautomatisierung. Er denkt dabei auch über Szenarien in der Wirtschaft und am Arbeitsmarkt nach, die eintreten könnten, falls sich die starke KI-These als richtig erweisen sollte und wir das Niveau einer *Artificial General Intelligence* erreichten, einer KI also, die in jeder Hinsicht der Intelligenz des Menschen mindestens ebenbürtig ist. Es könnten sich dann bewusst gewordene KI-Systeme und Maschinen von ihrem ‚Sklavendasein' emanzipieren und zu einem neuen KI-Proletariat werden oder auch den Menschen als Arbeitnehmer komplett verdrängen, was nach Weiß zu Massenarbeitslosigkeit und Verelendung führte.

Reinhard Kahle problematisiert in seinem Text „Wohin (ver)führt uns die neue KI?" den Umstand, dass KI immer häufiger Entscheidungen trifft vor dem Hintergrund, dass diese Entscheidungen nicht einem menschlichen Abwägungs- und Beurteilungsprozess, sondern lediglich einem automatisierten Blick auf Durchschnittswerte der Vergangenheit entspringen. Dies führt, so Kahle, zu einer problematischen

Konformität und dem Ausbleiben von Innovationen in Entscheidungsprozessen z.B. in Personalabteilungen. Des Weiteren betrachtet er die Herausforderungen und Gefahren etwa für unsere persönliche Freiheit im Zusammenhang mit selbstfahrenden Autos oder KI-Systemen, die zu Autoren der Medienwelt werden.

Auch **Uwe Engel** und **Holger Schultheis** beschäftigen sich mit Entscheidungen und gehen in ihrem Artikel „KI assistiert, der Mensch entscheidet. Ergebnisse der ersten Runde des Delphi-Surveys ‚Blick in die Zukunft. Wie Künstliche Intelligenz das Leben verändern wird'" auf Basis einer repräsentativen Bevölkerungsumfrage in Kombination mit einer Trend-Studie unter Wissenschaftlern der Frage nach, wer in der Zukunft in unterschiedlichen Szenarien nach Einschätzung der Befragten Entscheidungen treffen wird. Dabei wird zwischen den drei Optionen Künstliche Intelligenz, der Mensch und der Mensch mit Unterstützung von KI unterschieden. Die Ergebnisse der Studien werfen ein interessantes Licht auf die heutigen Erwartungen an die Zukunft der KI und werden zudem von den Autoren genutzt, um Fragen nach dem Vertrauen in KI-basierte Entscheidungen und moralische Aspekte zu diskutieren.

Die letzten drei Aufsätze in diesem zweiten Teil wenden sich explizit dem Feld solcher moralischen Fragen zu, die sich durch die Interaktion von Menschen mit KI-Systemen schon heute stellen und in der Zukunft noch relevanter werden. Wenn es zunehmend selbstverständlicher wird, dass Menschen gemeinsam mit KI-Systemen oder KI-gesteuerten Robotern interagieren, dann kann dies nicht ohne Auswirkungen auf unsere sozialen Beziehungen, das Handeln untereinander bleiben.

John Michael fragt in „Interaktionen mit Robotern – Wie verbindlich kann das sein?" nach der Verbindlichkeit bei gemeinsamen Handlungen mit KI-Systemen. Er geht davon aus, dass sich aus der sozialen Interaktion mit Robotern erhebliche Konsequenzen ergeben werden, da es zu einem Gemeinschafts- und Verbindlichkeitsgefühl gegenüber Maschinen und Computern kommen würde. Zugleich könnte es seiner Ansicht nach aber die Akzeptanz und Verbreitung von sozialen Robotern erleichtern, wenn ihnen gegenüber auch ein Gefühl der Verbindlichkeit aufgebaut werden kann, da es z.B. beim Einsatz von Pflegerobotern darauf ankommt, dass die von ihnen kommenden Anweisungen etwa zur Medikamenteneinnahme auch verbindlich befolgt werden.

Die sich daran anschließende Frage danach, welche Tiefe solche Mensch-KI-Beziehungen erreichen können, geht **Hendrik Kempt** in seinem Text „*Zwischenmenschlichkeit für Maschinen*" nach. Er glaubt, dass es gegenüber artifiziellen Systemen durchaus eine einfache Form

von Freundschaften geben könne, verneint aber die Möglichkeit von tieferen Beziehungen im Sinne aristotelischer Tugendfreundschaften, die für ihn grundsätzlich auf eine Mensch-Mensch-Beziehung hinauslaufen. Bei Robotern mangelt es dagegen an der gemeinsamen Basis der *Conditio Humana*, die durch Altern, Leid und Freude, Sterben, der Einzigartigkeit jeder Person und Körperlichkeit geprägt sei. Zudem weist Kempt auf die unabsehbaren Konsequenzen hin, die mit einer Aufnahme von Maschinen in einen Freundschaftskontext verbunden sind.

Im vorletzten Aufsatz des zweiten Teils „*Rechtfertigende Wachsamkeit gegenüber KI*" stellt **Ophelia Deroy** die Frage, ob man KI vertrauen kann. Sie bezweifelt das und rät uns dazu, wachsam zu bleiben. Es brauche zunächst eine Klärung, was unter KI verstanden wird und wie wir uns ihr gegenüber verhalten. Beides stehe in einem interdependenten Verhältnis zueinander, weil das, was KI für uns ist, immer auch abhängig ist von unseren Reaktionen auf sie und den anthropomorphistischen psychischen Zuständen, die wir ihr zuschreiben. Die Frage nach dem Vertrauen in KI könne deshalb nicht auf den Glauben an Verlässlichkeit verkürzt werden, sondern müsse erweitert werden um die Gründe für die Rechtfertigung von Vertrauen. Hier zeige sich jenseits einer strategischen der Mangel an einer moralischen und sozialen Rechtfertigung.

Schließlich führt uns der Aufsatz von **Catrin Misselhorn** „*Grundsätze der Maschinenethik*" in das spannende Feld dieser noch jungen Bereichsethik ein. Sie entwickelt dabei drei Grundsätze in Bezug auf die Frage, wie man „gute moralische Maschinen" bauen kann. Es gelte, 1. die Selbstbestimmung des Menschen zu fördern, statt einzuschränken, KI solle 2. nicht über Leben und Tod von Menschen entscheiden und 3. müsse für alle Aktionen von KI-Systemen gelten, dass immer der Mensch, und nicht die KI die Verantwortung trägt.

Künstliche Intelligenz mag Großartiges oder aber Beängstigendes verheißen, in jedem Fall regt sie zur Diskussion und philosophischen Auseinandersetzung mit ihr an, wofür der nun folgende zweite Teil unseres Bandes in spannender Weise Zeugnis ablegt.

Literatur

Aristoteles (1969): *Nikomachische Ethik.* Reclam, Stuttgart 1969.

Dennett, Daniel (1978): Intentional Systems. In: *Brainstorms. Philosophical Essays on Mind and Psychology.* Harvester Press, Hassocks 1978, S. 3-22.

Dennett, Daniel (1993): *Consciousness Explained*. Penguin Books, London 1993.

Dennett, Daniel (1995): *Darwin's Dangerous Idea – Evolution and the Meanings of Life.* Simon & Schuster Paperbacks, New York 1995.

Dennett, Daniel (1998): The Practical Requirements for Making a Conscious Robot. In: Ders.: *Brainchildren – Essays on Designing Minds.* The MIT Press, Cambridge 1998, S. 153-170.

Dennett, Daniel (2019): Wesen und Werkzeuge. In: *Süddeutsche Zeitung* vom 2.4.2019.

Dreyfus, Hubert (1972): *What Computers Can't Do Of Artificial Reason*. Harper & Row, New York 1972.

Duden Online: https://www.duden.de/rechtschreibung/Intelligenz, abgerufen am 9.8.2020.

Fingerhut, Joerg; Hufendiek, Rebekka; Wild, Markus [Hg.]: *Philosophie der Verkörperung.* Suhrkamp, Berlin 2013.

Heidegger, Martin (2006): *Sein und Zeit*. Max Niemeyer Verlag, Tübingen 2006.

Gödel, Kurt (1931): Über formal unentscheidbare Sätze der „Principia Mathematica" und verwandter Systeme. In: *Monatshefte für Mathematik und Physik* 38, 173-198.

Lucas, John Randolph (1961): Mind, Machines and Gödel. In: *Philosophy* 36 (137), 112-127.

Misselhorn, Catrin (2018): *Grundfragen der Maschinenethik*. Reclam, Stuttgart 2018.

Russell, Stuart; Norvig, Peter (2016): *Artificial Intelligence. A Modern Approach*. Pearson, Essex 2016.

Reese, Byron (2018): *The Fourth Age. Smart Robots, Conscious Computers, and the Future of Humanity*. Atria International, New York 2018.

Searle, John (1980): Minds, brains, and programs. In: *Behavioral and Brain Sciences* 3 (3), 417-424.

Tremmel, Sylvester (2020): KI lernt philosophieren. In: *heise online,* 8.8.2020, https://www.heise.de/news/KI-lernt-Philosophieren-4865882.html, abgerufen am 9.11. 2020.

Turing, Alan (1950): Computing machinery and intelligence. In: *Mind* 49, 433-460.