Referenz: Spiekermann, S. „Zum Unterschied zwischen künstlicher und menschlicher Intelligenz und den ethischen Implikationen der Verwechselung“, in: Ethische Herausforderungen im Zeitalter des Digitalen Wandels, Hrsg.: Österreichischer Rat für Forschung & Technologieentwicklung, ECOWIN, Wien, 2020

Zum Unterschied zwischen künstlicher und menschlicher Intelligenz und den ethischen Implikationen der Verwechselung

Sarah Spiekermann[1]

Einleitung: Fünf Thesen zu den Grenzen von KI-Systemen

Wenn Persönlichkeiten wie Bill Gates oder Elon Musk heute öffentlich vor KI-Systemen warnen, dann verknüpfen technische Laien diese Warnungen sehr schnell mit den Dystopien der Science-Fiction. Filme wie „Terminator“, „Blade Runner“ oder „Ex Machina“ haben die öffentliche Vorstellung davon geprägt, was KI-Systeme sein sollen: meistens ein Roboter mit menschenähnlicher Hülle und menschenähnlicher Intelligenz. Oder alternativ ein virtueller Softwareagent wie der Bordcomputer Hal in Stanley Kubricks „Odyssee 2001“. Nicht nur Laien teilen diese fast schon regulative Idee von unserer Zukunft, sondern auch viele Expert*innen glauben an das baldige Erscheinen solcher KI-Systeme, die zum einen erstaunlich humanoid sein sollen, zugleich aber auch viel intelligenter als wir Menschen (Bostrom, 2014; Kurzweil, 2006). Schenkt man dieser vermeintlichen IT-Zukunft unhinterfragten Glauben, dann kann man vor den skizzierten KI-Systemen schnell so viel Angst bekommen, wie der antike Mensch die Götterwelt fürchtete, die ähnlich anthropomorph gestaltet war. Dabei sollte jedoch nicht vergessen werden, dass Science-Fiction-Geschichten nichts mehr sind als moderne Legenden. Die kritische Auseinandersetzung mit Legenden hat schon immer die Geistesgeschichte geprägt und ist wichtig für kritisches Denken. Legenden sollten jedoch nicht die Grundlage bilden, welche die Politik, Wirtschaft und Wissenschaften im Technologiebereich dominiert.

Vor diesem Hintergrund werde ich im vorliegenden Beitrag herausarbeiten, dass KI-Systeme keineswegs so „heilig“, „heldenhaft“ oder „superintelligent“ sind, wie die Legenden nahelegen. Stattdessen möchte ich diese Systeme so beschreiben wie sie sind. Solch ein nüchterner

[1] Ich möchte an dieser Stelle Prof. Friedemann Mattern, Prof. Johannes Hoff und Jana Korunovska danken, die den vorliegenden Beitrag begutachtet und mit mir kritisch diskutiert haben.

Rahmen scheint mir überfällig, da politische Dokumente, wie etwa der National Defense Authorization Act (US Congress, 2018) KI-Systeme als „*menschenähnlich*“ beschreiben und sich damit erschreckend nah an die Science Fiction heran bewegen. Die Idee ist so mächtig geworden, dass offen über eigene Rechte für KI-Systeme nachgedacht wird („Robot Rights“; Gunkel, 2018), sogar schon Staatsbürgerschaften an diese vergeben worden sind (Hatmaker, 2017) und die Idee einer eigenen KI-Rechtspersönlichkeit die Bühne der Politik betreten hat (Krempl, 2018). Spricht man KI-Systemen eine solche Menschenähnlichkeit jedoch politisch zu und gibt man ihnen auf dieser Grundlage entsprechende Einsatzfelder und Rechte, dann greift man gleichzeitig zutiefst in die Freiheit und Würde von Menschen ein. In allen Bereichen des Lebens müssten sich neue Normen im Umgang mit der neuen „Spezies“ entwickeln. Ist es recht, einen Roboter zu treten? Tötet man ihn, wenn man ihn ausschaltet oder den Strom abdreht? Untergräbt der Fremdgang mit einem Roboter die Würde des menschlichen Partners? Die Palette an ethischen Fragen wäre schier unerschöpflich, ebenso wie die Fülle der sich daraus ergebenden Vorschriften. Dieser Beitrag stellt daher speziell die Frage in den Mittelpunkt, ob die Beschreibung von KI-Systemen als *menschenähnlich* zulässig ist. Zur Beurteilung bedarf es eines nüchternen Vergleichs zwischen den kognitiven Fähigkeiten von KI-Systemen und der menschlichen Intelligenz. Ich bezweifele nicht, dass KI-Systeme auch in Zukunft über eine weiter steigerbare Rechenleistung verfügen können und dass sie (unter der Annahme einer deutlich besseren Datenqualität als heute und einer kontext-sensitiven Konstruktion) logisch so mächtig werden könnten, dass sie nicht nur in engen Anwendungsbereichen nützlich sind, sondern tatsächlich die Intelligenz von uns Menschen bereichern. Allerdings werden KI-Systeme trotz dieser zu erwartenden Leistungssteigerung aus folgenden Gründen wahrscheinlich nie *menschenähnlich* sein:

1. KI-Systeme verfügen über wenig menschenähnliche Informationen
2. KI-Systemen können nicht menschenähnlich reagieren
3. KI-Systeme können nicht menschenähnlich denken
4. KI-Systeme haben keine menschenähnliche Motivation
5. KI-Systeme haben keine menschenähnliche Autonomie

Diesen Thesen liegt die Annahme zugrunde, dass Computersysteme auch weiterhin aus anorganischem Material gefertigt werden und digital arbeiten. [2]

[2] Mir ist bewusst, dass diese Annahme von derzeitigen Experimenten untergraben wird, die Software auf organischen Materialien zu implementieren (siehe z.B.: https://www.pnas.org/content/117/4/1853 oder https://royalsocietypublishing.org/doi/full/10.1098/rsif.2017.0937). Solche Experimente sind allerdings so sehr

Zum klassischen Unterschied von Vernunft und Intelligenz

Zunächst sei vorangestellt, dass ich auf der *klassischen* Unterscheidung zwischen Vernunft und Intelligenz aufbaue, die in unserer modernen Welt durcheinander geraten scheint[3]: Für die großen Philosophen der Antike bis ins späte Mittelalter galt: Die *Vernunft* eines Menschen ist die Fähigkeit, rationale Argumente, Daten und Fakten etc. zu sammeln, aufzunehmen und sachlogisch zu kombinieren (Griechisch: *dianoia*, Latein: *ratio*). Normale Menschen können lernen, vernünftig zu handeln; etwa Informationen zu sammeln, sich diese zu merken, wiederzugeben und zu kombinieren. Sie haben weitestgehend Kontrolle über diesen Vorgang. Im Gegensatz zu dieser Vernunft besteht *Intelligenz* oder Verstand (Griechisch: *noûs*, Latein: *intellectus*) darin, das Wichtige vom Unwichtigen zu trennen, zu abstrahieren, weiter zu denken oder schlichtweg zu sehen, worauf es in einer Angelegenheit ankommt oder nicht. Über diesen *intellectus* haben Menschen wenig Kontrolle: Man kann sich nicht zwingen etwas zu verstehen, was man nicht versteht.[4] Aus diesem Grund schreibt man Menschen unterschiedliche Grade einzigartiger Intelligenz zu.

Wie manifestiert sich nun Intelligenz im Gegensatz zur rationalen Vernunft? Erstens erlebt man sie körperlich: der Blutdruck steigt freudig im Augenblick des Geistesblitzes oder man genießt einen emotional angenehmen „flow", wenn die eigene Kompetenz mit den Anforderungen der Aufgabe perfekt zusammen fällt (Csikszentmihalyi, 1991). Zweiten ist es kaum möglich, Punkt-für-Punkt zu rekapitulieren, was es genau war, das einem aufgegangen ist oder was man wie in diesem Flow genau getan hat. Das wahre Verstehen, das dem intelligenten (nicht mechanischen) Handeln immanent ist, lässt sich nur eingeschränkt dingfest machen; geschweige denn in scharfen Informationseinheiten genau rekapitulieren.[5] Und drittens ist zu

---

in den Kinderschuhen, dass sie im Jahr 2020 wissenschaftlich kaum ernst zu nehmen sind; insbesondere, da die Kontrollierbarkeit der Reaktionen von organischen Materialen sich gar nicht mit den derzeitigen Paradigmen unserer Computermechanik und Statistik vereinbaren lässt: etwa der Berechenbarkeit, Nachvollziehbarkeit oder Wiederholbarkeit von Operationen.

[3] Vgl.:"If, in philosophy, there is a "*before* and *after* Immanuel Kant" (1724-1804), this is because he has inverted the meaning of *intelligence (Verstand)* and *reason (Vernunft)* as understood by all preceding philosophers: from Plato, Aristotle, Plotinus and St. Augustine to St. Thomas Aquinas, Dante, Leibniz, Malebranche, and beyond, all said to labor under an illusion which he alone was able to recognize and dispel! Indeed, in keeping with his conviction that intuition can only be sensible or empirical, he elevated *reason* to the highest rank among cognitive faculties, capable supposedly of rendering synthetic, systematic, universal and unified intelligibility. Hence intelligence or *intellect* came to be seen as inferior to reason: a secondary faculty concerned with processing abstractions, endowing sense experience with a conceptual form, and connecting the resultant concepts so as to constitute a coherent structure — until, finally, it turned into discursive knowledge, that is to say, *became* "reason." (Bérard, 2018)

[4] "We absolutely cannot think what we can't think" (G.E. Moore)

[5] Wenn überhaupt, schaffen es gute Lehrer, mit Hilfe von Analogien und Narrativen das klar zu machen, worum es geht. Solch eine Erklärung beginnt dann meistens mit den Worten: „Stell Dir vor…"

beachten, dass Intelligenz – im Gegensatz zur Vernunft – in der klassischen Lehre als „*noûs*“ bezeichnet wird; das gleiche Wort wie das französische „nous“, was zu Deutsch „wir“ heißt.[6] Die geteilte Wortwurzel weist darauf hin, dass Intelligenz etwas mit Gemeinschaftlichkeit zu tun hat; genauer gesagt: mit der Verbindung zu den Dingen, der Natur, den beteiligten Personen etc. Intelligenz setzt ein *geteiltes* Verständnis voraus.[7] Wenn man so ein geteiltes Verständnis nicht aufbauen kann, sagt man schnell: „Damit kann ich nichts anfangen!“ Man meint damit, dass man keine Intuition hat, wie man mit der gegebenen Sache umgehen soll.

Für das Verstehen des Unterschieds zwischen menschlicher und künstlicher Intelligenz scheint dieses „sich mit etwas verbinden“ zentral.[8] Es ist konstitutiv für den Unterschied zwischen Vernunft und Intelligenz: Die Vernunft zerlegt eine Sachlage völlig neutral in Einzelteile, analysiert diese nüchtern ohne körperliche Reaktion und trifft möglichst Punkt-für-Punkt nachvollziehbare Entscheidungen. Die Intelligenz hingegen erfordert, dass man sich mit einer Sache verbindet, dass man etwas mit ihr anfangen kann. Das Fazit dieses Artikels wird sein, dass KI-Systeme zwar vernünftig sein können, aber nicht intelligent. Und der Grund dafür ist letztlich, dass sie nicht die Fähigkeit haben, sich mit der Welt auf intelligente Weise zu verbinden; sie können mit der Welt „nichts anfangen“.

KI-Systeme verfügen über wenig menschenähnliche Informationen

Menschen sind hoch sensible, Körper-Geist-integrierte Systeme (Damasio, Everitt, & Bishop, 1996), die mit ihrer Umwelt permanent in Resonanz stehen (Rosa, 2016); so mächtige Systeme in der Tat, dass die Wissenschaft derzeit die These nicht mehr vermeiden kann, unser Gehirn als einen analogen Quantencomputer zu verstehen (Wendt, 2015). Selbst wenn ich mich dieser These hier nicht anschließe, so verarbeitet der Mensch in jedem Fall permanent optisch, akustisch, haptisch, gustatorisch und olfaktorisch Umweltinformationen. „Unsere Wahrnehmung … ist … das Produkt leibgebundener Formen *synästhetischer* Wahrnehmung“,

[6] https://en.wiktionary.org/wiki/nous#Etymology

[7] Deswegen ist es auch so angenehm, wenn man einen intelligenten Menschen sprechen hört, denn man anerkennt intuitiv sofort, dass der- oder diejenige richtig liegt. Meistens kann man nicht sagen, warum man meint, dass der intelligente Mensch richtig liegt, aber man teilt das Wirklichkeitsverständnis mit ihm oder ihr.

[8] Vgl.: „Unsere Wahrnehmung ... ist … das Produkt leibgebundener Formen *synästhetischer* Wahrnehmung. An ihrem Ausgangspunkt steht immer das, was die aristotelische Wahrnehmungslehre als *sensus communis* (Gemeingefühl) bezeichnete. Wir sehen ‚sprudelndes Wasser‘, hören ‚helle Glockentöne‘, sehen einen ‚harten Aufschlag‘, riechen ‚den stechenden Geruch von Heu‘, – und lernen erst später, ‚das Sprudelnde‘, ‚das Helle‘ oder ‚das Harte und Stechende‘ verschiedenen Sinnesmodalitäten zuzuordnen, die sich analytisch voneinander isolieren und vermeintlich ‚elementaren‘ (auditiven, visuellen, taktilen, olfaktorischen oder gustatorischen) ‚vorgegebenen Sinneseindrücken‘ zuordnen lassen.“ (Zitat, unveröffentlichtes Manuskript, Hoff, 2020)

schreibt Johannes Hoff. Und ohne mich hier der Argumentationskette eines Ray Kurzweil anschließen zu wollen (Feser, 2013; Kurzweil, 2006), so müssten *menschenähnliche* KI-Systeme zunächst einmal über ein sensomotorisch ähnlich mächtiges „Gehäuse" verfügen, dass auf ähnlich vollständige Weise die Umwelt verarbeitet. Davon ist der Stand der Technik jedoch sehr weit entfernt. Die mächtigsten der heutigen Supercomputer können gerade mal 1-2 Prozent der neuronalen Aktivität eines menschlichen Gehirns simulieren, wofür aber um ein Tausendfaches und mehr an Energie aufgebracht werde muss, als was unser biologisches System erfordert (Meier, 2017).[9] Ein Gehirn hat mehr als 860.000 Milliarden Neuronen, die sich zu über einer Trillion Synapsen miteinander verbinden. Früher dachte man, dass Menschen durch die Veränderung der Wirksamkeit bestehender Synapsen lernen. Diese Idee wurde auch zur Grundlage des maschinellen Lernens („Machine Learning"). Deshalb spricht man bei KI-System auch vom Trainieren von „Neuronen". Heute jedoch weiß man aus den Neurowissenschaften, dass es bei Menschen permanent zum Entstehen neuer Synapsen zwischen den Neuronen kommt. Mit anderen Worten: Das Gehirn verdrahtet sich permanent neu. Darf man technischen Quellen vertrauen, werden bis zu 40 Prozent der Synapsen auf einem Neuron täglich ersetzt. Ein ernüchterter Autor der KI-Welt schließt daher: "Während es wahr ist, dass die heutigen KI-Techniken auf die Neurowissenschaften verweisen, so verwenden sie dabei ein übermäßig vereinfachtes Neuronenmodell, das wesentliche Merkmale realer Neuronen auslässt; und die ML Neuronen sind auch auf eine Weise miteinander verbunden, die nicht die Realität der komplexen Architektur unseres Gehirns widerspiegeln" (S. 35 in Hawkins, 2017). Fakt ist also, dass die vielfältige, nuancierte und reiche menschliche Verarbeitung von Umweltinformationen unendlich feiner und flexibler ist als das, was wir von Maschinen in absehbarer Zeit und zu sinnvollen Energiekosten erwarten dürfen. Wenn jedoch die von KI-Systemen gesammelten Informationen andere sind, die noch dazu gröber verarbeitet werden, wie soll ein KI-System dann menschenähnlich sein oder (re)agieren? Weniger Daten, weniger Sensorium und weniger mächtige, dabei zudem völlig anders geartete Verarbeitung von Umweltinformation wird uns die KI-Systeme wahrscheinlich noch lange ‚ungelenk' vorkommen lassen.

[9] Vgl.: „As an example, consider a simulation that Markus Diesmann and his colleagues conducted several years ago using nearly 83.000 processors on the K supercomputer in Japan. Simulating 1.73 billion neurons consumed 10 billion times as much energy as an equivalent size portion of the brain, even though it used very simplified models and did not perform any learning…The IBM TrueNorth group, for example recently estimated that a synaptic transmission in its system costs 26 picojoules. Although this is about a thousand times the energy of the same action in a biological system, it is approaching 1/100.000 of the energy that would be consumed by a simulation carried out on a conventional general-purpose machine " (p. 29 and 31 in Meier, 2017).

Trotz dieser Sachlage haben KI-Systeme jedoch etwas, womit sie zu einer gewissen Vernunft in dieser Welt beitragen könnten. In ihre Gehäuse lassen sich nämlich eine Vielzahl von Sensoren einbauen, die wir Menschen wiederum nicht besitzen. Und diese Sensorinformation kann durch globale Vernetzung zwischen den Systemen effizient geteilt werden. Je nach Ausstattung können KI-Systeme beispielsweise Radioaktivität, Wärme oder Feuchtigkeit messen, die Anzahl der genutzten W-LANS in der Umgebung ermitteln, die Art der Möbel in umliegenden Häusern auslesen (etwa wenn diese mit RFID Chips auf bestimmten Frequenzbänder markiert sind). KI-Systeme wie mobile Roboter könnten (sofern das offiziell erlaubt werden sollte) über internationale Datenmärkte das sozio-demographische Profil jedes vorbeigehenden Passanten beziehen oder durch optische Analyse der Gesichtsmimik relativ treffsicher auf deren momentane Gefühlslage schließen. Auch wenn KI-Systeme sich also mangels körperlichen Sensoriums und mangels menschenähnlicher Gehirn- bzw. Körper-Geist-Strukturen nicht in gleicher Weise mit der lebendigen, sinnlichen Welt verbinden können, so können sie sich doch untereinander vernetzen und ungeheuer viele anders geartete Daten aggregieren. Das Resultat ist eine ganz eigene Maschinenratio, deren Bedeutung für das menschliche Leben und Wirtschaften wichtig sein kann, auch wenn bzw. gerade weil sie nicht menschenähnlich ist.

KI-Systeme können nicht menschenähnlich reagieren

Wenn Menschen auf ihre Umwelt reagieren, dann zeigt sich darin in ganz besonderer Weise ihre Intelligenz im Sinne des „*noûs*“. Jeder, der eine normal ausgeprägte Verbindung zur Welt aufbauen kann, reagiert auf diese Welt unmittelbar und gefühlsmäßig, was in allen Lebenslagen wichtig ist. Man nehme etwa eine moralisch aufgeladene Situation, in der man ein Unrecht beobachtet. Im Angesicht negativ konnotierter Zustände wie Unrecht, Gemeinheit, Brutalität oder ähnlichem kennt unsere Sprache Ausdrücke für unser sofort einsetzendes Wertgefühl; etwa dass man angesichts des Beobachteten „Bauchschmerzen hat“, dass es einem „kalt über den Rücken läuft“ oder dass es einem „die Haare zu Berge sträubt“. Aber nicht nur in negativen Situationen leitet unser Wertgefühl unser Verstehen. Das meiste, was unserem Leben in Form von Werten letztlich Bedeutung gibt – ist Sympathie, Liebe, Freundschaft, Gemeinschaft, Sicherheit –, erschließt sich uns Menschen dadurch, dass wir uns gefühlsmäßig hingezogen fühlen (Scheler, 1921,2007). Das Sich-hingezogen-fühlen oder Abgestoßen-sein bestimmt dann oft maßgeblich unsere Reaktion.

Es genügt wohl der Hinweis, dass Computersysteme aller Generationen und auch kein KI-Systeme über einen Leib verfügen, der solche oder ähnliche Empfindungen haben kann. Ein Gehäuse aus Stahl hat keine Bauchschmerzen und es läuft ihm auch nicht kalt den Rücken runter.[10] Ein humanoider Roboter könnte zwar sagen, dass er Bauchschmerzen hat. Die Reaktion der Maschine besteht dann in dieser Sprachausgabe. Diese Simulation einer Menschenähnlichkeit macht das KI-System jedoch nicht wirklich menschenähnlich.

Was das Beispiel mit den simulierten Bauchschmerzen darüber hinaus interessant macht, ist, dass wahre Menschenähnlichkeit oft gar nicht gewünscht wird. Epley, Waytz und Cacioppo (2007) zeigen, dass viele Menschen lediglich aufgrund von Gefühlen der Einsamkeit zu Anthropomorphismen neigen. Wäre es nicht schön, einen völlig neutralen Freund zu haben, der gleichzeitig hoch emphatisch reagiert bzw. in seiner Simulation von Empathie uns gegenüber so vollkommen ist, dass man sich in der Interaktion geborgen fühlt? Empathisch-wirkende, aber in Wirklichkeit emotionslose Vernunft scheint das Wunschbild einer Zeit, in der unser Vertrauen in Menschen selbst am Tiefpunkt ist.[11]

Dieses mangelnde Vertrauen wird der menschlichen Natur meines Erachtens nicht gerecht: Beim Menschen korreliert die Empathiefähigkeit mit der Aktivität von Spiegelneuronen, so dass es zu einer natürlichen Empathiefähigkeit kommt (Jenson & Iacoboni, 2011). KI-Systeme hingegen haben keine Spiegelneuronen. Was KI-Systeme mittelfristig allerdings mehr nutzen könnten, sind Sensoren, welche in der Lage zum Erkennen menschlicher Emotionen und Reaktionen sind. Es ist Computern schon heute möglich, minimale Gesichtsregungen, Pupillenerweiterung oder Hautreaktionen präzise zu messen und daraus relativ verlässlich abzuleiten, wie sich ein Mensch gerade fühlt. Ferner können KI-Systeme die technischen Skills besitzen, auf solche Beobachtungen des menschlichen Gegenüber vernünftig zu reagieren.[12] Ich gebrauche hier bewusst den Begriff der Vernunft, denn das Computersystem ist ja nicht in der Lage, ein „*noûs*“ zu aktualisieren und sich mit dem Gegenüber emotional zu verbinden. Es kann jedoch ein rationales Modell des menschlichen Gegenübers errechnen und dann bestimmte

[10] Es ist mir bewusst, dass einige Wissenschaftler wie Daniel Dennett argumentieren, dass der Mangel an Resonanzfähigkeit eines Systems nicht an dessen Materialeigenschaften liegt. Ein Beweis für diese Argumentation liegt jedoch nicht vor. Fakt ist, dass Computer in absehbarer Zeit nicht über einen resonanzsensiblen Leib verfügen.

[11] Vergleiche dazu meinen Artikel zum schlechten Menschenbild unserer Zeit (Spiekermann, 2019a) und den historischen Quellen dieses Denkens (Spiekermann, 2019b).

[12] Man beachte an dieser Stelle, dass die technischen „skills“ von KI, also technische Komponenten, die bestimmte Algorithmen ausführen, zu unterscheiden sind von dem was Richard Sennet unter „skill“ versteht (Sennett, 2009)

prädeterminierte oder erlernte Reaktionen auf diesen Menschen ausführen. Die Frage ist, wie man solch eine Maschinenreaktion beurteilt und ob man sie als „intelligent“ im menschenähnlichen Sinne gelten lassen will.

Was einen Menschen emotional intelligent macht ist, dass er oder sie eins mit sich und der Welt ist. Er oder sie versteht und kennt sich selbst und ist aus dieser Selbsterkenntnis heraus, gepaart mit einem Verstehen seiner Umwelt, in der Lage, auf diese Umwelt in einem verbindenden Sinne zu reagieren. Beim KI-System fehlt neben der Fähigkeit, sich wahrhaft zu verbinden, noch ein zweiter Aspekt: Dass es eben dieses menschliche Selbstbewusstsein nicht hat. Es hat kein Selbst, was es in die Verbindung mit seinem menschlichen Gegenüber einbringen könnte, und damit ist jede Reaktion sinnbildlich „selbstlos“. Das aber ist nicht menschlich.[13]

KI-Systeme können nicht menschenähnlich denken

Wenn Computersysteme „denken“, dann ist das, was sie eigentlich tun, rechnen. Jedes Computersystem, inklusive aller Formen von KI, basiert auf Daten, die nach einem ganz spezifischen Schema kodiert, eingelesen, in Datenbanken klassifiziert, strukturiert, funktional integriert, idealerweise mit Metadaten beschrieben und ggf. als Teil von Ontologien abgespeichert sind. Das, was häufig als KI-System-spezifische Funktionalität beschrieben wird, etwa das maschinelle Lernen (z.B. mit „Deep Neural Networks“) ist Teil von genau dieser Datenverarbeitungsarchitektur. Sie erlaubt, dass Rohdaten nicht nur als Information abgespeichert werden, sondern sinnvoll „repräsentiert“ werden und dass diese Repräsentationen sich auch verändern können. KI-Systeme können Muster erkennen und anpassen, zum Beispiel in unserer Sprache. Und sie können dann in Kombination mit dem Wissen der Linguistik gewachsene Modelle aufbauen (synthetisieren), was ihnen erlaubt, Sprechaktionen zu erkennen, zu analysieren und selbst auszuführen.

Gerade bei der KI-unterstützen Datenverarbeitung kann diese Aggregation kontinuierlich sein. Abgespeicherte Informationen und Repräsentationen verändern sich permanent, mit neu

[13] Es sei an dieser Stelle gesagt, dass es Momente in der Interaktion mit Robotern gibt, wo diese gerade aufgrund ihrer selbstlosen Reaktionen ungeheuer verwundbar wirken und dadurch wiederum menschlich (Vgl. dazu Spiekermann, 2019b). Und wichtig ist mir auch anzumerken, dass ich natürlich keinem Menschen absprechen möchte, dass er oder sie oftmals selbstlos bzw. altruistisch handelt. Nur ist es normalerweise schon so, dass wir unser Selbst auch in selbstlose Handlungen einbringen. Und selbst bei altruistischen Formen des Handelns spielt diese eigene Psyche und Motivlage eine Rolle.

einfließenden Datensätzen. Beobachtet man große Datensätze am Bildschirm *life* bei einer solchen Verarbeitung, dann kann man den Eindruck gewinnen, dass dieser Fluss von Daten und sich verändernden Informationsobjekten eine eigene Lebendigkeit entfaltet. Eine Lebendigkeit, die ich als „synthetische Existenz" bezeichne. Diese synthetische Existenz ist beeindruckend, wenn man sie am Werk sieht. „Es blinkt, es lebt" (Gehring 2004) mag die beeindruckte Beobachter*in eines derartigen Systems rufen. Allerdings darf man bei allem Enthusiasmus diese blinkenden Visualisierungen nicht mit einer *menschenähnlich*en Existenz gleichsetzen. Es ist nicht das Leben selbst, sondern lediglich eine darauf aufsetzende, beobachtende Instanz der realen Phänomene. Beobachter und Beobachtetes fallen logisch nicht zusammen.

Der Unterschied zwischen menschlichem Denken und künstlicher Informationsverarbeitung ist, dass Menschen in der Regel keine Daten in einem Modell zusammenrechnen. Es ist zwar durchaus so, dass wir in der Psychologie und in der Ökonomie eine lange Tradition haben, menschliches Entscheiden und Handel so zu modellieren. So nutzen wir – leider – noch immer die Vorstellung vom „Homo Oeconomicus", um den Menschen als eine Art „Präferenzoptimierer" bzw. „Nutzenmaximierer" abzubilden. Auch in der Psychologie verwenden wir Modelle, wie beispielsweise jene der unterschiedlichen Theorien vom Vernünftigen Handeln[14], um zu erklären, wie Menschen agieren. Die Anzahl der Modelle, die menschliches Denken im Rahmen von Entscheidungsprozessen beschreiben, ist groß. Aber jeder vernünftige Wissenschaftler weiß auch, dass all diese Modelle, die menschliches Denken bezüglich Entscheidungen in Einzelteile zerlegen und in der abhängigen Variable des Handelns (oder intendierten Handelns) wieder zusammensetzen, nur grobe Heuristiken des tatsächlichen menschlichen Denkens und Handelns abbilden. Das macht die Modelle nicht weniger wertvoll. Heuristiken sind wissenschaftlich wichtig, um uns selbst als Spezies besser zu verstehen. Aber sie sind nicht geeignet, das menschliche Denken und Handeln an sich vollständig abzubilden oder verlässlich vorherzusagen. Wer das nicht akzeptieren will, mag demütig an die Bestimmtheitsmaße und Fehlergrößen erinnert werden, die mit jeder wissenschaftlichen Statistik einhergehen. Nur äußerst selten zerlegt ein Mensch in einer Entscheidungssituation minuziös die Aspekte in Einzelbestandteile bzw. startet von diesen ausgehend und rechnet sie dann mit Gewichtungen wieder zusammen; auch wenn uns so mancher Verhaltensökonom nahelegen mag, dies öfter mal zu tun.

[14] Siehe etwa „Theory of Reasoned Action" oder „Theoy of Planned Behavior (vlg. Ajzen &Fishbein 2005).

Stattdessen scheint gesichert, dass Menschen ihre Umwelt normalerweise in nicht-summativen, ganzheitlichen Gestalten wahrnehmen und erinnern; zumindest dann, wenn beide Gehirnhälften gesund zusammenarbeiten (Mc Gilchrist, 2009). Die rechte Gehirnhälfte, die für diese ganzheitliche Wahrnehmung zuständig ist, interagiert dabei mit der linken Gehirnhälfte, die das Wahrgenommene (zum Beispiel durch das Sprachzentrum) strukturiert (Mc Gilchrist, 2009). Bereits Anfang des 20. Jahrhunderts nutzte Edmund Husserl das antike Konzept der „*noemata*", um die ganzheitlichen Gestalten unseres Denkens begrifflich zu fassen (Husserl, 1993).
Noemata erlauben uns Menschen, die ‚Sinngestalt' eines Phänomens zu erfassen, und wir tun dies nicht, indem wir einzelne Datenpunkte zu dieser Gestalt irgendwie rechnerisch gewichtet synthetisieren, sondern diese geht uns auf oder wird uns intuitiv einsichtig. In Anlehnung an dieses Wissens spricht man in den Neurowissenschaften und der Gedächtnisforschung heute daher auch vom „autonoetischem Bewusstsein", wenn man die menschliche Persönlichkeit und ihr Gedächtnis beschreibt (Baddeley, Eysenck, & Anderson, 2015).[15] In weiten Teilen unseres Denkens aktualisieren wir Menschen das, was wir beobachten, in *noemata*, für die wir Begriffe gefunden haben.

So könnte man sagen, dass ein Mensch zum Beispiel eine Idee davon teilen kann, was es heißt, *gut* zu sein[16]. Wenn es dann in der Umgebung zu einem Vorfall kommt, wo sich jemand dieser Vorstellung bzw. Idee entsprechend verhält, dann erkennt ein Mensch das sehr schnell. Ein KI-System hingegen hat keine geteilte Idee vom Guten. Es kann trainiert werden, eine ganz bestimmte Abfolge von Handlungen als etwas zu erkennen, was als „gut" oder „richtig" gekennzeichnet (zu Englisch: „label") worden ist und daher die (gelernte oder determinierte) Regel inkorporieren, dass ein guter Mensch an einer roten Ampel stehen bleibt. Aber es erkennt dann immer nur diese eine Ausprägung des Guten wieder. Läuft also jemand über die rote Ampel, weil er ein Kind retten will, wird das KI-System errechnen, dass das nicht gut ist; es sei denn, es hat genau diese Sequenz vorher schon einmal gelernt (bzw. mit anderen verteilten KI-Systemen geteilt). Menschen hingegen sind sofort in der Lage, die Idee des Guten in der rettenden Tat zu erkennen. Ein KI-System folgt einer bottom-up „pattern recognition theory of mind" (Feser, 2013). Der Mensch hingegen nutzt top-down die Wiedererkennung von *noemata*, die sich noch dazu gar nicht in Datenpunkten fassen lassen, sondern nur in Form von ganzheitlichen Wesensgestalten aufscheinen. Diese Dynamik des Denkens erlaubt unserer

[15] Der autonoetische Teil des Gedächtnisses ist Teil des Langzeitgedächtnisses und derjenige Teil, welcher die gewachsene Persönlichkeit eines Menschen widerspiegelt.
[16] Wiederum weist uns das Wort *noema* mit der Wortwurzel des *nous* auf das Geteilte hin; das, was von der Gemeinschaft geteilt verstanden wird.

Spezies problemlos mit der unstrukturierten Vielzahl sensomotorischer, optischer, olfaktorischer und haptischer Umwelteinflüsse klarzukommen, die keinerlei Kompilierung bzw. Übersetzung in Dateneinheiten, Vorverarbeitung, „Training", vordefinierte Datenbankfelder, Ontologien, etc. bedürfen.

Mir ist natürlich bewusst, dass so manchem modernen Wissenschaftler (unterschiedlicher Disziplinen) nicht wohl dabei ist, diese Phänomenologie anzuerkennen. Denken in *noemata*? Keine messbaren und fein abgrenzbaren Informationseinheiten? Das erregt Unwohlsein. Dies ist nicht verwunderlich, denn der dominierende Teil der Wissenschaftler*innen hängt nach wie vor dem an, was Johannes Hoff als „Baukästchenmetaphysik" beschreibt. Er schreibt: „So wie Johannes Gutenberg aus beweglichen Lettern seine Druckformen ‚zusammensetzte', so ‚synthetisiert' das kantische ‚Ich' aus ‚mannigfaltigen' Sinneseindrücken wahrnehmbare Objekte."[17] Nach Humes ist sogar das synthetisierende ‚Ich' auf ein „Bündel unterschiedlicher Zustände"[18] reduzierbar (Hoff, 2020). Dies scheint jedoch ein auslaufendes Modell vom menschlichen Denken, denn folgt man den aktuellsten neurowissenschaftlichen Erkenntnissen, die eingebettete sind in eine *intelligente* geisteswissenschaftliche Tradition, dann wird folgendes klar: „Wenn es überhaupt als ‚Maschine' bezeichnet werden kann, dann ist das Gehirn keine ‚Synthetisierungs-' oder ‚Projektionsmaschine', sondern eine ‚Inhibitions-'oder ‚Selektionsmaschine'. Das ‚Geistige' wird durch das Gehirn nicht *generiert*, sondern im Zusammenspiel mit anderen Organen und korrespondierenden Umwelteindrücken, die den Möglichkeitsraum des Wiss- und Wahrnehmbaren einschränken, auf mehr oder weniger diskrete Formgestalten *kontrahiert."*[19]

Der fundamentale Unterschied zwischen einem maschinellen, also synthetisierenden und einem menschlichen, d.h. kontrahierenden Denken hat ethische Implikationen. KI-Systeme können nur Daten verarbeiten, für die sie eine entsprechende technische Repräsentanz haben („trainiert" wurden). So lange ein KI-System keine Repräsentanz für alle erdenklichen, ja selbst die unwahrscheinlichsten Situationen vorhält, wird es immer wieder die unsinnigsten Fehler

[17] Nach Kant ist Erkenntnis ein „Ganzes verglichener und verknüpfter Vorstellungen" (Kant, *Kritik der reinen Vernunft*, A 97). An ihrem Ausgangspunkt steht eine in den Sinnen passiv und zerstreut gegebene Mannigfaltigkeit. Zu deren Synthesis ist die „Spontaneität unseres Denkens" erforderlich (A77 / B102). Ein Objekt ist folgerichtig „das, in dessen Begriff das Mannigfaltige einer gegebenen Anschauung *vereinigt* ist" (B137). Erst dadurch werden sachbezogene Urteile möglich, die uns erlauben, die Welt zu erkennen, indem sie den referentiellen Gegenstandsbezug subjektiver Synthesen sichern.

[18] Roth (2001), 338.

[19] Hierzu unter Bezugnahme auf Aristotles' Wahrnehmungstheorie: Fuchs (2016), 187f sowie Aristoteles, *De Anima / Über die Seele*, III, 430f.

machen. Es wird falsche Urteile fällen, sobald es mit einer Situation konfrontiert ist, die im „Training“ nicht vorkam. Das jedoch ist fatal, weil das ganze menschliche Leben nun mal eine Abfolge kontext-sensibler, nicht-identischer Wiederholungen ist.[20]

Weil KI-Systeme Fehler machen, die für die betroffenen Menschen mehr als unangenehm, sogar gefährlich werden können, gibt es bis heute fast ausschließlich sog. „enge KIs“ (Englisch: „narrow AIs“). Diese KI-Systeme werden für einen ganz bestimmten, geschlossenen Kontext trainiert. Hier können sie die theoretisch möglichen Datenmuster erlernen und sogar Details erkennen und Eventualitäten antizipieren, die ein Mensch oft gar nicht erkennt. Das zeigt aber noch einmal, warum ein KI-System überhaupt nicht *menschenähnlich* ist. Es leidet am sog. „underfitting“ im offenen, generellen, nicht-identisch wiederholten Lebenskontext, den Menschen und Gruppen teilen. Und es ist präziser und vorausschauender als ein Mensch in geschlossenen Kontexten sich wiederholender Gegebenheiten.

Wenn man diesen Unterschied zwischen menschlichem Denken und der Datenverarbeitung von KI-Systemen nicht versteht, kann es zu einem ethisch problematischen Einsatz der Technik kommen – und zwar überall dort, wo die komplexe individuelle Lebenssituation eines Menschen mit Hilfe eines KI-Systems erfasst werden soll; etwa bei der Frage, ob jemand ein Krimineller ist und bleibt, eine Straftat begehen wird oder begangen hat, an einer Universität oder in einem Job gute Leistung bringen kann, etc. Die nicht-identischen Charakterzüge von einzelnen Menschen in ihren nicht identischen Lebenssituationen, die sie in nicht identischen Kontexten durchleben, sind so einzigartig, dass es einem KI-System nicht möglich ist, sie zu erfassen. KI-Systeme, die auf solch generellem Lebensterrain eingesetzt werden, laufen permanent in Gefahr des *underfitting*.

KI-Systeme haben keine menschenähnliche Motivation

In der Science-Fiction werden KI-Systeme immer dann spannend, wenn sie sich ihre eigenen Ziele setzen; etwa der Bordcomputer Hal im Stanley Kubricks Film Odysee 2001. Wenn Menschen sich bewusst Ziele setzen, dann deswegen, weil ihnen eine Handlung, ein Handlungsergebnis oder eine Seinsweise als sinnvoll bzw. wertvoll erscheint. Oftmals sind es

[20] Kierkegaard, S. (2005). *Die Krankheit zum Tode – Furcht und Zittern – Die Wiederholung – Der Begriff der Angst.* München: DTV.

aber gar nicht bestimmte Ziele, sondern eher Werte, die eine Art Zugqualität auf Menschen ausüben und sie dazu motivieren, in dieser oder jener Form zu handeln.[21] Seit vielen Jahrzehnten setzt sich die Motivations- und Verhaltensforschung mit diesen Mechanismen im Detail auseinander. Hier spricht man von „Motiven", die bei Menschen generell, kontextuell oder situations-spezifisch geprägt sind (Vallerand, 1997). So kann ein Mensch eine generelle Neigung zu einem bestimmten Verhalten haben; McClelland unterscheidet etwa Menschen mit einer relativ ausgeprägten Tendenz zu Macht, Leistung oder Zugehörigkeit (McClelland, 2009). Oder es werden in wiederkehrenden Kontexten (Freizeit, Familie, Lernen) immer wieder bestimmte intrinsische Motive wirksam; zum Beispiel die Neugierde, der Wunsch nach Ordnung oder ein gewisser Idealismus (Reiss, 2004). Ebenso gibt es ganz situationsspezifische Motive: zum Beispiel gewinnen zu wollen oder den Wunsch, in Ruhe gelassen zu werden. Grundsätzlich geht man in der Psychologie davon aus, dass es solche Motive sind, die das Verhalten von Menschen prägen.

Angesichts dieser menschlichen Motive stellt sich die Frage, wie man einem KI-System so emotional aufgeladene Werte wie Macht, Zugehörigkeit, Leistungsfreude oder Idealismus beibringen soll. KI-Systeme haben weder den gedanklichen Zugang zu diesen Begriffen in Form von *noemata*, noch haben sie ein Leibgedächtnis, welches ihnen die sinnliche Wertigkeit dieser Motive vermitteln könnte. Und selbst der beste Theoretiker könnte ein Motiv wie Macht oder Bedürfnis nach Ruhe nicht präzise modellieren.

Wo sich Psychologie und KI treffen, ist wiederum in vereinfachenden Modellen der menschlichen Ratio, etwa der Erwartungswerttheorie. Die Erwartungswerttheorie postuliert, dass Menschen in eine Art Kalkül eintreten, wo sie sich ausrechnen, ob ein bestimmtes Verhalten zur Erfüllung des angestrebten Motivs beiträgt (Vroom, 1964). Erwartungswertfunktionen können eine spannende Grundlage sein, um eine KI-System zu optimieren. Allerdings ist es einem KI-System nur möglich, sog. „extrinsische" Motive abzubilden. Zum Beispiel eine Geldmenge, die durch ein Verhalten erzielt wird. KI-Systeme an Finanzmärkten werden mit solchen monetären Zielfunktionen trainiert. Sobald man allerdings von Einsatzkontexten absieht, wo eine einfache Maximierungslogik Sinn macht, und sich in den Normalfall menschlicher Lebensbereiche begibt, wo Motive um ihrer selbst willen angestrebt werden, dann läuft sich das KI-System tot, weil es zu solchen Motiven keinen Zugang hat (siehe oben).

[21] Siehe dazu etwa Scheler (1921, 2007).

Irreführend ist, wenn sich einige KI-Experten dennoch auf den schlüpfrigen Pfad begeben, KI-Systemen eine menschenähnliche Motivation zuschreiben. Sie verwenden dazu Begriffe, etwa den der „intrinsischen Belohnung". Wenn man sich dann jedoch genauer anschaut, wie sie diese definieren bzw. im System modellieren, dann weicht diese „intrinsische Motivation" doch ganz erheblich von dem ab, was die Psychologie darunter versteht. Intrinsische Motivation von KI-Systemen ist bei Schmidhuber etwa das, was eine Reinforcement Learning Komponente in einem KI-System übernimmt, deren mathematische Zielfunktion sich bei neu entdeckten Datenmustern maximiert und dann aus diesem Kalkül heraus weitere Aktionen im System anstößt (Schmidhuber, 2010).[22] Intrinsische Motivation im menschlichen Sinne ist jedoch nicht primär an etwas Neues gebunden: Ganz im Gegenteil handelt es sich bei intrinsischer Motivation um das (wert)schätzende Erleben von etwas nicht-identisch Wiederholtem oder innerlich ersehnt Angenehmen.[23] . So sind zum Beispiel Zustände wie „Zugehörigkeit zu empfinden" oder „in Ruhe zu sein" solche die nicht identisch wiederholt werden (können), sondern etwas schon Erlebtes darstellen, etwas, das man (schon) kennt, aber doch immer wieder neu erfährt.[24] Ferner hat die intrinsische Motivation auch nichts mit Maximierung zu tun, so wie das bei einer „Reinforcement Komponente" der Fall ist, sondern ist ganz im Gegenteil ein Handeln um des Motivs selber willen, ein Handeln, dessen Maximum gar nicht angestrebt wird, sondern eher dessen Stimmigkeit. Kurz: Das Entlehnen des psychologischen Begriffs durch Informatiker wie Schmidhuber (2010) ist irreführend. Intrinsische Motivation im menschlichen Sinne ist bei KI-Systemen nicht herzustellen.

KI-Systeme haben keine sozial eingebettete Autonomie

Der letzte große Bereich, der die angebliche Menschenähnlichkeit von KI-Systemen rechtfertigen soll, ist der ihrer potenziellen Autonomie. Das US-amerikanische Defense Science Board definiert die technische Autonomie eines KI- Systems als „die Fähigkeit [des KI-Systems], unabhängig aus verschiedenen Vorgehensweisen auszusuchen bzw. diese zusammenzustellen, um ein Ziel zu erreichen, was auf dem Wissen, dem Weltverständnis, dem Selbstverständnis und der Situation aufbaut" (Summer Report) Diese technische Autonomie

[22] Ein bisher nicht da gewesenes Datenmuster wird als ‚neu' klassifiziert. Dieses ‚neu' ist automatisch gut. Die „Reinforcement Learning Komponente" maximiert sich und „belohnt" das darunter liegende System.
[23] Beachte: „intrinsisch" steht für „aus dem Inneren kommend".
[24] Lediglich die primitive Neugierde ist ein Motiv (von vielen), wo das Neue in Reinform gut sein mag.

setzt ein, wenn das System einmal aktiviert ist. Wird beispielsweise eine Drohne auf Mission geschickt, kann sie so konfiguriert werden, dass sie dann auf Basis selbst gesetzter Ziele agiert. Oder ein Stromnetz kann mit Hilfe von Smart-Meter-Daten die Netzstabilität selbständig managen. Mit dieser Definition hat sich das Militärgremium für die höchsten Grade dessen entschieden, was es „Autonomie" nennt: Die gesamte Kontrolle liegt bei der Maschine (Parasuraman & Sheridan, 2000). Notwendig ist an dieser Stelle an die Argumentation der Kantianer*innen zu erinnern, wonach man nur einen Sklaven, der eben gerade nicht autonom ist, „auf Mission" schicken kann, da es ihm nicht möglich ist, sich selbst auf Mission zu schicken, nicht selbst die Art der Mission wählen kann und die Mission auch nicht ablehnen kann. Kantianer würden die Art von Mission, die eine Drohne erfüllt, als „Heteronomie" bezeichnen. Wiederum benutzt die Informatik also einen philosophisch sehr präzise definierten Begriff und schreibt ihren Maschinen damit Fähigkeiten zu, die diese nicht besitzen. Ausgenommen sind natürlich jene KI-Systeme aus der Science-Fiction, die sich tatsächlich vollkommen eigenständig ihre Missionen aussuchen und ihre Ziele stecken. Das sind sog. „Generelle KIs", die sich nach einem nicht-überwachten Machine-Learning-Verfahren selbst Ziele setzen können sollen – KIs, die es in sinnvoller Weise derzeit nicht nur nicht gibt, sondern von denen auch nicht klar ist, ob es sie jemals geben wird oder geben sollte!

Doch selbst wenn es eines Tages Systeme solcher „Generellen KIs" geben sollte, sind diese dennoch nicht menschenähnlich. Warum nicht? In der sog. „Self-Determination Theory" haben Ryan and Deci (aufbauend auf der Motivationsforschung der 1970er Jahre) seit 2000 immer wieder nachgewiesen, wie wichtig für Menschen die drei Faktoren der eigenen Kompetenz, der Autonomie und der menschlichen Bezogenheit sind (Ryan & Deci, 2000). Autonomie wird hier verstanden als die Möglichkeit, für das eigene Handeln ursächlich sein zu können und zwar so, dass man in Harmonie mit sich selbst agieren kann; dass man sich etwa nicht von Fremdeinflüssen gezwungen fühlt, bestimmte Handlungen anzustoßen. Das bedeutet allerdings bezogen auf uns Menschen *nicht*, dass man im Ausleben dieser Autonomie komplett frei ist von den Wünschen, Zielen und Gewohnheiten der Gruppe, mit der man sich verbunden fühlt (Deci & Vansteenkiste, 2004). Ganz im Gegenteil: Der Mensch ist ein *zoon politicon*, ein soziales Wesen. Das heißt, dass die vernünftige Entscheidung eines Menschen in der Regel das soziale Umfeld mitberücksichtigt. Menschen leben, wenn überhaupt, „autonom" nur in einer Art sozial eingebetteter Autonomie. Ihre Freiheit endet dort, wo die Freiheit der anderen

beginnt.[25] Wenn man sich dieses Spannungsfeld von eigener Freiheit und menschlich „autonomen" Entscheidungen anschaut, was innerhalb eines sozialen Umfelds stattfindet, dann erkennt man sehr bald, dass es die Verletzbarkeit des Menschen ist, die ganz wesentlich dazu beiträgt, dass er oder sie sich aus sich selbst heraus dazu entscheidet, oft für Andere mitzudenken und nicht nur als frei schwebendes Individuum autonom – also frei von äußeren Einflüssen – zu entscheiden.[26] Die ganze aristotelische Tugendlehre beschäftigt sich mit dieser Thematik des Menschen, ein gesundes Maß zu halten und in der Art seiner Entscheidungen weder durch ein Zuviel noch ein Zuwenig negativ in seine Gruppe hinein zu wirken. Was ihn oder sie jedoch dazu motivieren kann, dieses Maß zu halten, ist die Verletzbarkeit; etwa von der eigenen Gruppe nicht anerkannt zu werden, ausgestoßen zu werden, allein zu sein.

An genau dieser Stelle unterscheidet sich die gelebte menschliche „Autonomie" ganz maßgeblich von der „Autonomie" eines KI-Systems. Zweitere nämlich wird vor allem als Möglichkeit verstanden, nach eigenem Kalkül eine Aktion anzustoßen, ohne eine Bestätigung vom Operator einzuholen. Soziale Erwägungen spielen dabei für die Maschine keine Rolle, denn sie ist nicht verletzlich. Es ist ihr egal, ob sie keinen Strom mehr bekommt oder abgeschossen wird, denn die Idee des Todes ist ihr im menschlichen Sinne nicht vermittelbar.[27]

Zu einer achtsamen Definition von KI-Systemen und deren Abgrenzung vom Menschen

Die Diskussion betreffend die vermeintliche Menschenähnlichkeit von KI-Systemen hat sich an einer Reihe von Charakteristika orientiert, anhand derer man diese Computersysteme definieren kann: ihren physischen Körpern, den Daten und ihrer Verarbeitung, den Quellen der Zielsetzung und der Autonomie. Abbildung 1 fasst große Teile des Beschriebenen zusammen. In der linken Spalte wird wiederholt, welche Ausprägungen KI-Systeme in unseren modernen Legenden besitzen, also in der Science-Fiction. In der mittleren und rechten Spalte, die für diesen Beitrag relevant sind, werden die KI-Systeme beschrieben, die in der Praxis existieren

[25] Vgl.: "To be autonomous does not mean to be detached from or independent of others, and in fact Ryan and Lynch (1989) showed how autonomy can be positively associated with relatedness and well-being. Autonomy involves being volitional, acting from one's integrated sense fo self, and endorsing one's actin. It does not entail being separate from, not relying upon, or being independent of others."

[26] Vgl. hierzu Spiekermann. (2020).

[27] Ein KI-System könnte natürlich eine Funktion integrieren, die das Abgeschaltet-werden oder das Abgeschossen-werden minimiert. Sie wird dann Handlungsstrategien entwickeln, die solche Eventualitäten vermeidet. Damit wird die Autonomie der Maschine begrenzt; allerdings nicht *sozial* begrenzt, wie das beim Menschen der Fall ist.

oder mit denen zumindest ernsthaft experimentiert wird. Die grau dargestellten Systemeigenschaften sind solche, die noch einen hohen Forschungsbedarf haben und bisher nicht verlässlich funktionieren; etwa die Verarbeitung von unstrukturierten Daten. Die jeweilige Ausgestaltung von Körpern, Daten, Lernweisen, Zielen und Graden der Autonomie bestimmen, welche kognitiven Prozesse ein KI-System jeweils durchführen kann. Daher werden diese Charakteristika unterhalb der KI-Kognition dargestellt. Ferner unterscheidet man, ob ein KI-System eine Software hinter einem technisch autonomen (eigentlich lediglich heteronomen) physischen System ist; ob man sich also auf eine rein virtuell erzeugte Entität bezieht oder ob man an ein Hardwaresystem denkt, welches die beschriebenen Charakteristika integriert. Für beide Formen von Systemen gibt es Praxisbeispiele. Rein virtuelle KI-Systeme sind zum Beispiel digitale Sprachassistenten wie Amazons „Alexa“ oder der „Google Speech Assistant“. Im Gegensatz dazu haben physische Systeme, wie etwa selbstfahrende Autos, Aktuatoren, die die errechneten Handlungen eines Algorithmus für ein System in mechanische Bewegungen übersetzen. In jedem Fall empfiehlt es sich, immer von einem „KI-System“ zu sprechen, weil in der Regel eine Vielzahl von Algorithmen verknüpft werden, ergänzt um die entsprechenden Datenbanken und die ausführenden (motorischen oder virtuellen) Systemelemente. Ein KI-System wirkt in seiner Gesamtheit auf den Menschen oft *intelligent*. Es hängt aber von dieser Einschätzung nicht ab, um sich als KI-System zu qualifizieren.

Vor diesem Hintergrund definiere ich ein KI-System hier als *ein virtuelles und/oder physisches, integriertes Computersystem, welches auf Basis von zumindest teilweise unstrukturierten und reichhaltigen Datensätzen diverse kognitive Funktionen selbständig ausführen kann und unter Berücksichtig von i.d.R. von Menschen beeinflussten Zielfunktionen in der Lage ist, auch ohne menschliche Intervention wirksame Aktionen durchzuführen. Kognitive Funktionen können das Wahrnehmen, Planen, Schlussfolgern, Kommunizieren und Entscheiden einschließen.*

Nimmt man nun alle bis hierher erläuterten Bereiche zusammen, in denen sich KI-Systeme von Menschen doch maßgeblich unterscheiden, so muss man sich fragen, wie Expert*innen bis in die höchsten politischen Ebenen hinauf auf die Idee kommen können, KI-Systemen ernsthaft Menschenähnlichkeit zuzuschreiben. So haben Menschen beispielsweise zwar identische DNA-Basispaarketten mit anderen Säugetieren – diese Ketten stimmen angeblich zwischen Menschen und Schweinen zu 90 % überein –, dennoch würde niemand auf die Idee kommen, Menschen mit Schweinen zu verwechseln. Und niemand würde auf die Idee kommen, Schweinen ähnliche Rechte wie Menschen zu geben.

Aus kritischer ethischer Perspektive stellt sich daher die Frage, ob es in Ordnung ist, Menschen mit KI-Systemen zu vergleichen oder ob dies nicht einer Diffamierung gleichkommt. Die von Marketing und Hypes geprägte Technikwelt setzt sich zu wenig achtsam mit der angestammten Bedeutung solcher Begriffe auseinander und begibt sich damit auf eine fragwürdige Gradwanderung, für die die Autoren Hastak und Mazis den Begriff der „stillschweigenden Täuschung“ geprägt haben (Hastak & Mazis, 2011). Aus Sicht der fröhlichen Wissenschaften könnte ein Mensch freilich spielerisch kontern: „Scharf und milde, grob und fein, vertraut und seltsam, schmutzig und rein, der Narren und Weisen Stelldichein: Dies alles bin ich, will ich sein, Taube zugleich, Schlange und Schwein“ (Nietzsche, 1882).

	Menschenähnliche Sci-Fi KI	Virtuelles KI System	Physisches KI-System
KOGNITIVE FUNKTIONEN	wahrnehmen (Informationssammlung- und integration), planen (z.B. mH.v. Musteranalyse), schlussfolgern (z.B. basierend auf Mustern), kommunizieren (z.B. text to voice), entscheiden/wählen (z.B. basierend auf prediction)		
KÖRPER	Menschenähnlicher Körper	Software (inkl. Datenbanken)	Software (inkl. Datenbanken) Hardware (inkl. Aktuatoren)
INFORMATION	Unbereinigte Daten noema	Unbereinigte Daten **Bereinigte Daten (strukturiert, kontinuierliche/diskrete)**	Unbereinigte Daten **Bereinigt (strukturiert, kontinuierliche/diskrete)**
MACHINE LEARNING (ML)	Unüberwachtes ML Minimale Intervention von Menschen im Life Cyle	Unüberwachtes ML **Überwachtes ML am Anfang, dann eingefroren** **Überwachtes ML während des Life Cycles**	Unüberwachtes ML **Überwachtes ML am Anfang, dann eingefroren** **Überwachtes ML während des Life Cycles**
ZIELE	Ziel beeinflusst vom Menschen Ziel selbst gegeben (nicht zufällig) Ziel jenseits des vorspezifizierten Einsatzbereichs	**Ziel gegeben vom Menschen** Ziel selbst gegeben (auch mH.v. Zufallsalgorithmen) Ziel jenseits des vorspezifizierten Einsatzbereichs	**Ziel gegeben vom Menschen**
AUTONOMIE (nach Aktivierung)	Agiert ohne Intervention vom Menschen	Agiert ohne Intervention vom Menschen Agiert mit Intervention vom Menschen	Agiert ohne Intervention vom Menschen Agiert mit Intervention vom Menschen

Abbildung 1: Charakteristika von realistischen und unrealistischen KI-System-Systemen

Literatur

Ajzen, I. &Fishbein, M. (2005). *The Influence of Attitudes on Behavior.* Mahwah, New Jersey, USA: Erlbaum.

Baddeley, A., Eysenck, M. W., & Anderson, M. C. (2015). *Memory*. New York: Psychology Press.

Bérard, B. (2018). *Unmasking „AI“*. Blog. https://philos-sophia.org/unmasking-ai/

Bostrom, N. (2014). *Superintelligenz: Szenarien einer kommenden Revolution*. Berlin: Suhrkamp Verlag.

Csikszentmihalyi, M. (1991). *Flow: The Psychology of Optimal Experience*. New York: Harper Collins.

Damasio, A. R., Everitt, B. J., & Bishop, D. (1996). The somatic marker hypothesis and the possible functions of the prefrontal cortex. *Philosophical Transactions: Biological Sciences, 351*(1346), 1413-1420.

Deci, E. L., & Vansteenkiste, M. (2004). Self-determination theory and basic need satisfaction: Understanding human development in positive psychology. *Ricerche di Psichologia* (27), 17-34.

Epley, N., Waytz, A., & Cacioppo, J. T. (2007). On Seeing Human: A Three-Factor Theory of Anthropomorphism. *Psychological Review,* 114(4), 864-886.

Feser, E. (2013). Kurzweil's Phantasms - A Review of How to Create a Mind: The Secret of Human Thought Revealed. *First Things*. https://www.firstthings.com/article/2013/04/kurzweils-phantasms

Fuchs, T. (2016). *Das Gehirn - ein Beziehungsorgan: Eine phänomenologisch-ökologische Konzeption* (5. Auflage ed.). Stuttgart: Kohlhammer.

Gehring, R. (2004). Es blinkt, es denkt. Die bildgebenden und die weltbildgebenden Verfahren der Neurowissenschaften. *Philosophische Rundschau,* 51, 272-295.

Gunkel, D. J. (2018). *Robot Rights*. Cambridge, US: MIT Press.

Hastak, M., & Mazis, M. B. (2011). Deception by Implication: A Typology of Truthful but Misleading Advertising and Labeling Claims. *Journal of Public Policy & Marketing,* 30(2), 157–167.

Hatmaker, T. (2017, 27.10.2017). Saudi Arabia bestows citizenship on a robot named Sophia. *Tech Crunch*. https://techcrunch.com/2017/10/26/saudi-arabia-robot-citizen-sophia/

Hawkins, J. (2017). What Intelligent Machines Need to Learn From the Neocortex. *IEEE SPECTRUM,* 54(6), 33-37.

Husserl, E. (1993). *Ideen zu einer reinen Phänomenologie und phänomenologischen Philosophie* (5. Auflage ed.). De Gruyter.

Jenson, D., & Iacoboni, M. (2011). Literary Biomimesis: Mirror Neurons and the Ontological Priority of Representation. *California Italian Studies,* 2(1). http://www.neurohumanitiestudies.eu/archive/paper/?id=150

John S. McCain National Defense Authorization Act for Fiscal Year 2019 (2018).

Krempl, S. (2018, 12.04.2018). Streit über "Persönlichkeitsstatus" von Robotern kocht hoch *heise online*. https://www.heise.de/newsticker/meldung/Streit-ueber-Persoenlichkeitsstatus-von-Robotern-kocht-hoch-4022256.html

Kurzweil, R. (2006). *The Singularity is Near- When Humans Transcend Biology*. London: Penguin Group.

Mc Gilchrist, I. (2009). *The Master and his Emissary - The Divided Brain and the Making of the Western World*. New Haven and London: Yale University Press.

McClelland, D. (2009). *Human Motivation*. Cambridge, UK: Cambridge University Press.

Meier, K. (2017). The Brain as Computer - The Brain May be Bad At Crunching Numbers, but it's a Marvel of Computational Efficiency. *IEEE Spektrum, 54*(6), 27-31.
Nietzsche, F. (1882). *Die Fröhliche Wissenschaft.* : Reclam.
Parasuraman, R., & Sheridan, T. B. (2000). A Model for Types and Levels of Human Interaction with Automation. *IEEE Transactions on Systems, Man, and Cybernetics,* 30(3), 286-297.
Reiss, S. (2004). Multifaceted nature of intrinsic motivation: The theory of 16 basic desires. *Review of General Psychology,* 8(3), 179-193.
Rosa, H. (2016). *Resonanz. Eine Soziologie der Weltbeziehung* (2nd Edition ed.). Berlin: Suhrkamp Verlag.
Roth, G. (2001). *Fühlen, Denken, Handeln. wie das Gehirn unser Verhalten steuert.* Frankfurt/Main: Suhrkamp Verlag.
Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist,* 55, 68-78. https://dx.doi.org/10.1037/0003-066X.55.1.68
Scheler, M. (1921, 2007). *Der Formalismus in der Ethik und die Materiale Wertethik - Neuer Versuch der Grundlegung eines ethischen Personalismus* (2. unveränderte Auflage ed.). Halle an der der Saale: Verlag Max Niemeyer.
Schmidhuber, J. (2010). Formal Theory of Creativity, Fun, Intrinsic Motivation (1990 - 2010). *IEEE Transactions on Autonomous Mental Development,* 2(3), 230-247.
Sennett, R. (2009). *The Craftsman*. New York: Penguin Books.
Spiekermann, S. (2019a, 23./24. März 2019). Der Mensch als Fehler. *Süddeutsche Zeitung,* 15.
Spiekermann, S. (2019b). *Digitale Ethik - Ein Wertesystem für das 21. Jahrhundert*. München: Droemer.
Spiekermann, S. (2020). Digitale Ethik und Künstliche Intelligenz. In: *Philosophisches Handbuch Künstliche Intelligenz.* Hrsg. Mainzer, K. München: Springer Verlag. (Im Erscheinen).
Vallerand, R. J. (1997). Toward a hierarchical model of intrinsic and extrinsic motivation. *Advances in Experimental Social Psychology,* 29, 271-360.
Vroom, V. H. (1964). *Work and Motivation*. New Jersey, USA: John Wiley & Sons Inc.
Wendt, A. (2015). *Quantum Mind and Social Science - Unifying physical and social ontology*. Cambridge UK: Cambridge University Press.