Erscheint in: Handbuch Künstliche Intelligenz und die Künste, Hrsg. von Stephanie Catani und Jasmin Pfeiffer, Reihe "De Gruyter Reference", De Gruyter Verlag Berlin, 2021.

Preprint Version

Zum Begriff der Künstlichen Intelligenz

Jana Koehler

"To create an artificial being has been the dream of man since the birth of science. Not merely the beginning of the modern age when our for-bearers astonished the world with the first thinking machines - primitive monsters that could play chess. How far we have come. The artificial being is a reality of perfect simulacrum, articulated in limb, articulated in speech, and not lacking in human response […]. But what does it amount to?"

William Hurt, amerikanischer Schauspieler, als Professor Hobby im Prolog des Films Artificial Intelligence von Steven Spielberg 200 (Ausschnitt aus Minute 1:06-2:34)

Eine grundlegende Frage nach den Grenzen der Berechenbarkeit und ein Name für ein Forschungsgebiet

1950 stellt der englische Mathematiker Alan Turing die Frage, ob menschliche Intelligenz durch einen Computer berechnet werden kann und schlägt auch gleich einen Test vor, mit dem die Intelligenz eines Computers im Wettbewerb mit einem Menschen testbar wird (Turing 1950).

"The new form of the problem can be described in terms of a game which we call the 'imitation game.' It is played with three people, a man (A), a woman (B), and an interrogator (C) who may be of either sex. The interrogator stays in a room apart front the other two. The object of the game for the interrogator is to determine which of the other two is the man and which is the woman […] It is A's object in the game to try and cause C to make the wrong identification. The object of the game for (B) is to help the interrogator. […] We now ask the question, ´What will happen when a machine takes the part of A in this game?' […] These questions replace our original, 'Can machines think?' […] The new problem has the advantage of drawing a fairly sharp line between the physical and the intellectual capacities of a man." (Turing 1950, 433 - 434)

Schaut man sich diese Originalformulierung von Alan Turing an, fällt der Genderaspekt des Tests ins Auge, der in späteren Interpretationen des Tests eher in Vergessenheit gerät und von Historikern mit der Verfolgung und Diskriminierung von Alan Turing aufgrund seiner Homosexualität erklärt wird. Auch wird im englischen Original nicht wirklich deutlich, ob ein Mann oder der Mensch allgemein gemeint ist, wenn Turing davon spricht, dass die Testformulierung den Vorteil hat, *„a fairly sharp line between the physical and the intellectual capacities of a man"* zu ziehen. (Turing 1950, 434)

1991 wurde der Turing Test in Form der Loebner Competition erstmalig durchgeführt und danach mehrere Male wiederholt (Epstein 1992). Die anhaltende Kritik, insbesondere auch von

Forschern, siehe zum Beispiel (Hayes und Ford 1995), zeigt jedoch die Problemhaftigkeit eines solchen Ansatzes auf und hat den Test in der Bedeutungslosigkeit verschwinden lassen. Die Frage, was eine Künstliche Intelligenz (KI) letztendlich aber ausmacht und ob und wie sie vom Menschen abgegrenzt werden soll, hat jedoch mit den jüngsten Fortschritten des Gebiets eine erneute Aufmerksamkeit erfahren und umfassende Diskussionen initiiert. Diese Aufmerksamkeit wird mit dem Sieg des Watson Systems im Spiel Jeopardy im Jahre 2011, bei dem Fragen zu von den Spielern gewählten Sachgebieten beantwortet müssen, noch weiter befeuert und so wundert es auch nicht, dass doch auch wieder der Turing Test in die Diskussion gerät (Shah 2011).

1956, nur wenige Jahre nach dem Erscheinen von Turings Artikel, treffen sich amerikanische Forscher am Dartmouth College an der amerikanischen Ostküste, um ein Forschungsprogramm zu initiieren, das der Frage nach einer intelligenten Maschine nachgehen soll (McCarthy et al. 1955). John McCarthy, einer der beteiligten Forscher, prägt den Namen für dieses Programm: Künstliche Intelligenz. 1996 nahm die Autorin dieses Beitrags an einer Podiumsdiskussion mit John McCarthy auf der International Conference on AI Planning (AIPS) in Edinburgh teil, an der McCarthy sich rückblickend eher nicht so glücklich mit der Namenswahl zeigte und die damalige Diskussion zu einem geeigneten Namen für das neue Forschungsgebiet so reflektierte: „Wenn ich zurückschaue und sehe, welche Auswirkungen dieser Name auf unser Gebiet hatte, denke ich, ich hätte diesen Namen vielleicht besser nicht vorschlagen sollen. Wir diskutierten damals zwei Kandidaten: Komplexe Computeranwendungen und Künstliche Intelligenz. Niemand fand, dass der erste Name dem neuen Gebiet gerecht wird und man muss auch verstehen, dass wir jung waren und einen spannenden Namen für unser Gebiet wollten.“ (Gedächtnisprotokoll der Autorin).

Fussnote: Siehe auch das Interview mit John McCarthy von 2007 auf youtube:

https://www.youtube.com/watch?v=KuU82i3hi8c

Im Vorschlag für das Forschungsprogramm zur Künstlichen Intelligenz von 1955 (McCarthy et al. 1955, 2) findet sich eine konkrete Liste an Forschungsfragen, die untersucht werden sollen und deren Lösbarkeit auf der Grundlage der Hypothese angenommen wird, dass jeder Aspekt eines lernenden und intelligenten Systems so präzise mathematisch beschreibbar ist, dass er von einem Computer simuliert werden kann.

“The study is to proceed on the basis of the conjecture that any aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it. An attempt will be made to find how to make machines use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves.” (McCarthy et al. 1955, 2)

Die im Zitat formulierte Hypothese wird später von den KI-Forschern Alan Newell und Herbert Simon in Form der sogenannten Physical Symbol System Hypothesis pointiert auf den Punkt gebracht und damit Intelligenz zunächst als reine Symbolverarbeitung positioniert.

"A physical symbol system has the necessary and sufficient means for general intelligent action." (Newell and Simon 1976,116)

In der KI-Forschung finden sich aber auch viele Vertreter anderer Positionen, die zum Bespiel die Notwendigkeit eines Körpers und die Auseinandersetzung mit der Umwelt als grundlegend

für ein im wahrsten Sinne Begreifen der Welt und damit auch intelligentes Verhalten ansehen, siehe u.a. (Brooks 1991).

Auch der nachfolgende Beitrag von Bianca Westermann beschäftigt sich intensiv mit der Frage nach der Verkörperung eines intelligenten Systems.

Die Suche nach der Definition der Künstlichen Intelligenz

So umfassend und vielfältig wie das Forschungsgebiet, so unterschiedlich sind auch die Versuche, Intelligenz bzw. Künstliche Intelligenz zu definieren. Eine Auflistung von über 70 Definitionen gibt der Bericht von Legg und Hutter (2007). Dabei fällt auf, dass für viele KI-Forscher, die Fähigkeit, eines Systems, selbstgestellte Ziele zu erreichen, als sehr entscheidend angesehen wird. In der Kunst, aber auch in der KI ist die Frage, inwieweit eine Maschine sich selbst Ziele stellen kann oder ob diese Ziele vom Menschen vorgegeben werden sollen, immer wieder Gegenstand zahlreicher Kontroversen. Besonders im Film werden oft Szenarien zu Ende gedacht, in denen sich von einer intelligenten Maschine gewählte Ziele gegen den Menschen wenden, denken wir zum Beispiel an Kubrick´s 2001: A Space Odyssey.

Für die KI-Forschung der letzten 30 Jahre ist insbesondere die Metapher des rationalen Agenten prägend, wie sie im Lehrbuch von Russell und Norvig (2010) formuliert ist. Danach ist ein Agent rational, wenn er die folgenden vier Bedingungen erfüllt:

1. Der Agent kann die Umwelt wahrnehmen.
2. Diese Wahrnehmungen dienen als Grundlage für die Entscheidungen, die der Agent trifft.
3. Die Entscheidungen führen zu Aktionen, die der Agent in der Umwelt ausführt.
4. Die Entscheidungen müssen rational sein, das heißt, sie müssen zur bestmöglichen Aktion führen, die der Agent als Reaktion auf die Wahrnehmungen in der Umwelt ausführen kann.

Dabei wird rationales Entscheiden (Denken) und Handeln als verschieden von menschlichem Denken und Handeln positioniert, da Menschen zum Beispiel auch altruistisch handeln, anstatt den persönlichen Nutzen maximierend.

Ausgehend von dieser Definition kann die Frage nach der Definition Künstlicher Intelligenz entlang von 4 Dimensionen beantwortet werden (Russell und Norvig 2003):

1. Geht es um rationales Denken, das logisch und optimierend abläuft?
2. Geht es um rationales Handeln auf der Grundlage bestmöglicher Aktionen, wobei rationales Denken nicht unbedingt eine Voraussetzung für rationales Handeln sein muss?
3. Soll KI menschliches Denken simulieren, mit all seinen Fehlern?
4. Soll KI menschliches Handeln imitieren, in seiner ganzen Vielfalt und so dem Menschen ebenbürtig sein in allen Aufgaben?

Die Auffassung und Haltung zu diesen Fragen bestimmen die Ausrichtung einzelner Forschungsgebiete innerhalb der KI. Der Unterschied zwischen einem Computer, der intelligent ist, und einem Computer, der menschliche Intelligenz imitiert, wird dabei als sehr grundlegend empfunden. Für einen Computer, der intelligent sein soll, aber menschliches

Denken und Handeln nicht nachbildet, können ganz andere Methoden und Ansätze verfolgt werden und ein Studium der menschlichen Intelligenz ist nicht erforderlich. Tatsächlich hat dieser rationale Ansatz die Fortschritte der KI-Forschung durch eine konsequente Ausrichtung an mathematischen Methoden stark geprägt.

Die nachfolgende Abbildung zeigt eine Auswahl wichtiger aktueller Forschungsgebiete und Methoden in der KI und ordnet sie den Teilgebieten der Mathematik zu, auf denen diese Forschungsgebiete und Methoden wesentlich (aber nicht ausschließlich) beruhen.

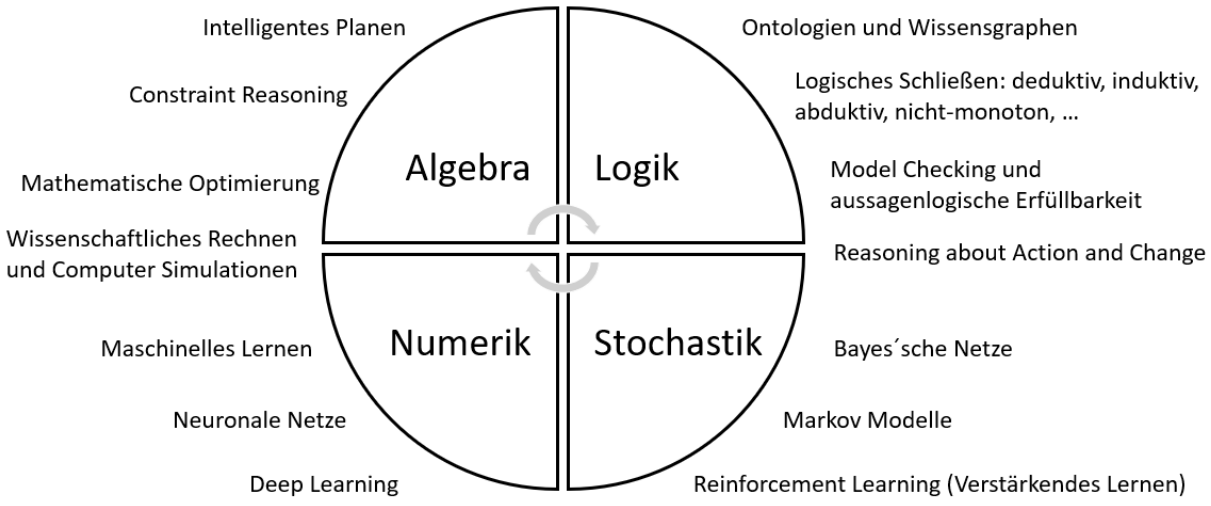

Abbildung 1: Überblick über wichtige Methoden und Gebiete der Künstlichen Intelligenz sowie ihre mathematischen Grundlagengebiete

Einige dieser Gebiete, wie zum Beispiel das Wissenschaftliche Rechnen, das u.a. mit Hilfe von Computersimulationen Fragestellungen aus allen wissenschaftlichen Disziplinen beantworten kann, wurden zwar im ursprünglichen Vorschlag von 1955 bereits angedacht, hatten sich aber in der Vergangenheit als eigenständige wissenschaftliche Gebiete entwickelt. Erst in jüngster Zeit ergibt sich wieder eine stärkere Verbindung zu den Methoden der KI.

„If a machine can do a job, then an automatic calculator can be programmed to simulate the machine. The speeds and memory capacities of present computers may be insufficient to simulate many of the higher functions of the human brain, but the major obstacle is not lack of machine capacity, but our inability to write programs taking full advantage of what we have." (McCarthy et al. 1955, 2)

Andere Gebiete, wie die mathematische Optimierung und Statistik sind in der Mathematik angesiedelt, stellen aber grundlegende Methoden zur Verfügung, die für die KI gerade in jüngster Zeit sehr wichtig sind. Insbesondere haben sich innerhalb der letzten 20 Jahre in der KI-Forschung stark datengetriebene Methoden entwickelt, für die statistische und quantitative Methoden zur Entscheidungsunterstützung unabdingbar sind. Aus historischer Sicht und basierend auf dem Ansatz des rationalen Denkens waren Methoden der mathematischen Logik in der KI vor allem in den 1980er und 1990er Jahren sehr einflussreich, um zum Beispiel logische Kalküle als Grundlage rationaler Denkgesetze einzusetzen (Wright 1996). In dieser Zeitspanne wurde in der KI sehr intensiv untersucht, inwieweit menschliches Wissen in mathematischer Logik, insbesondere der sehr einfachen Aussagenlogik oder auch in Teilen der

Prädikatenlogik, dargestellt werden kann, und inwieweit menschliches Denken sich durch logische Regeln abbilden lässt.

Die Unschärfe menschlichen Denkens und die Komplexität der realen Welt hat einer rein Logik-basierten KI jedoch sehr schnell die Grenzen aufgezeigt. Einerseits wurden logische Kalküle oft zu komplex und waren damit auch für Berechnungen schwer handhabbar, wie z.B. Deontische oder Nichtmonotone Logiken, andererseits ließen sich für viele Probleme nicht wirklich überzeugende und in allen Situationen funktionierende Lösungen finden. Sehr einfach lässt sich dies am Beispiel des sogenannten Nixon Diamanten (eigentlich der Nixon-Raute) verdeutlichen, einem berühmten Beispiel aus der Geschichte der KI. Mit den folgenden 4 Aussagen

a) *„Nixon ist ein Quäker.“*
b) *„Nixon ist ein Republikaner.“*
c) *„Quäker sind Pazifisten.*
d) *„Republikaner sind keine Pazifisten.“*

lässt sich mit Hilfe der klassischen Logik aus a) und c) ableiten, dass Nixon Pazifist ist, während aus b) und d) folgt, dass Nixon kein Pazifist ist, womit ein logischer Widerspruch entsteht.

Die Schwierigkeit, subtile semantische Unterschiede zu modellieren, lässt sich anhand des folgenden Beispiels leicht erkennen. Mit dem Wissen *„Max ist ein Mensch“* und *„kein Mensch kann fliegen“*, lässt sich mit Modus Ponens korrekt schlussfolgern, dass Max nicht fliegen kann. Erfahren wir nun aber, dass Max Pilot ist, dann ist dieser Schluss nicht mehr korrekt, denn Max kann sehr wohl fliegen, aber hier ist mit dem Verb *fliegen* eine andere Art von Fliegen, nämlich das Fliegen mit einem Flugzeug, gemeint.

Berühmt ist auch das von John McCarthy formulierte Frame Problem (McCarthy und Hayes 1969), bei dem es darum geht, all das zu notieren bzw. zu berechnen, was sich in der Umwelt NICHT ändert, wenn ein Agent eine Aktion ausführt. Dies Problem ist bis heute nicht zufriedenstellend gelöst. Betrachten wir einen Roboter, der eine Schachtel greift. Alle Objekte in der Schachtel werden in der Schachtel bleiben, wenn der Roboter die Schachtel transportiert. Ebenso werden sich andere Objekte in der Umgebung des Roboters nicht bewegen, zum Beispiel das Regal, von dem der Roboter die Schachtel genommen hat. Oder doch nicht? Was passiert, wenn der Roboter die Schachtel leicht schräg hält oder wenn er sich abrupt dreht. Werden Objekte aus der schräg gehaltenen Schachtel fallen und wird der Roboter das Regal umstoßen? Keine Regel ohne beliebig viele Ausnahmen, keine Aktion, deren direkte oder indirekte Effekte sich wirklich vollständig beschreiben ließen.

Fußnote: Siehe auch https://plato.stanford.edu/entries/frame-problem/ für eine sehr schöne Diskussion zum aktuellen Stand und den philosophischen Fragestellungen im Zusammenhang mit dem Frame Problem.

Mit der Entwicklung der datengetriebenen KI, die einerseits von den durch die Digitalisierung erzeugten umfangreichen und diversen Datenmengen, aber auch von einer immer verfügbaren hohen Rechenleistung in Form des Cloud Computing über das Internet profitiert, rückten in den letzten Jahren vor allem die Methoden des Maschinellen Lernens in den Mittelpunkt (Russell und Norvig, 2020). Dabei werden drei Formen von Lernen unterschieden:

1. Beim Nicht-überwachten (oder auch unüberwachten) Lernen werden Algorithmen eingesetzt, die nach Mustern und häufig auftretenden Zusammenhängen in Daten

suchen und Datensätze anhand ihrer Ähnlichkeit gruppieren. Dabei benötigen diese Algorithmen keine Beispiele, wie sie gruppieren sollen oder welche Daten welche Muster aufweisen, anhand derer sie zunächst trainiert werden müssen. Sie kommen immer dann zum Einsatz, wenn unklar ist, worin Ähnlichkeiten zwischen Daten bestehen könnten, da sie diese sehr gut aufspüren. Zum Beispiel kann ein solches Verfahren in einem Datensatz von Kunden diejenigen Gruppen identifizieren, die ähnliche Kaufgewohnheiten haben. Ebenso sind die Verfahren sehr gut geeignet, Anomalien und Unregelmäßigkeiten in Daten aufzuspüren.

2. Das überwachte Lernen benötigt Trainingsdaten, in denen Daten als Paare von Eingabewerten und gewünschten Ausgabewerten vorliegen. Anhand dieser Daten können überwachte Lernalgorithmen die Funktion approximieren, die zu einem gegebenen Eingabewert den richtigen Ausgabewert berechnet. Wird ein solcher Algorithmus zum Beispiel mit Bildern trainiert, die mit diversen Eingabeparametern beschrieben werden und jeweils als Ausgabewert das Objekt annotieren, das auf dem Bild dargestellt ist, dann können diese Algorithmen sehr zuverlässig in der Bilderkennung eingesetzt werden, um das dargestellte Objekt zu identifizieren.

3. Das verstärkende Lernen erlaubt es Algorithmen durch aktives Experimentieren zu lernen. Dabei werden Aktionen in einer echten oder simulierten Umgebung ausgeführt, um ein konkretes vorgegebenes Ziel zu erreichen. Je nach Erfolg oder Misserfolg einer Aktion erhalten diese Algorithmen eine positive oder negative Rückmeldung aus der Umgebung, die ihnen hilft, die Erfolgsaussichten einer Aktion in einer bestimmten Situation besser einzuschätzen und zukünftig möglichst erfolgversprechende Aktionen zu bevorzugen. Ein autonomes Auto kann so zum Beispiel lernen, einzuparken, in dem es immer wieder versucht, eine Parklücke ohne Kollisionen zu erreichen. An diesem Beispiel wird sofort deutlich, dass das Lernen eher in einer simulierten Umgebung erfolgen sollte, da in der realen Umgebung zu viele Risiken vorhanden sind und fehlschlagende Aktionen echte Schäden verursachen können.

Zu den wichtigsten Algorithmen im überwachten Lernen gehören die Neuronalen Netze, für die die Grundlagen bereits in den 1940er Jahren gelegt wurden (McCulloch und Pitts, 1943). Neuronale Netze sind Graphen, d.h. sie bestehen aus Knoten, die mit Kanten untereinander verbunden sind. Abbildung 2 zeigt links eine mögliche Anordnung der Knoten und Kanten eines Neuronalen Netzes. Ganz links (blaue Kästchen) die Eingabeschicht. In der Mitte folgend 5 verborgene Schichten, rechts die Ausgabeschicht, hier nur bestehend aus einem einzelnen Knoten (rosa Kästchen). Ein Eingabeobjekt, z.B. ein Bild, wird durch eine Vielzahl von Parametern, z.B. die Farbe eines bestimmten Pixels im Bild, beschrieben. Ein Knoten der Eingabeschicht des Netzes steht dabei für genau einen Parameter („Pixel 5“) und nimmt den Wert dieses Parameters 0,437 als Eingabe auf. Dieser Wert stellt den RGB Code 112130056 für die Farbe „Olivengrün“ dar, der auf eine Zahl zwischen 0 und 1 normalisiert wurde. Zusätzlich enthalten die Eingabeschicht und die verborgenen Schichten jeweils noch einen zusätzlichen Knoten („Bias Neuron“), dessen Wert 1 ist und der es Neuronalen Netzen erlaubt zum Beispiel auch dann einen Ausgabewert zu berechnen, wenn alle Eingabewerte 0 sind. Die Kanten eines Neuronalen Netzes sind mit Gewichten (in der Regel Werte zwischen -1 und 1) versehen, die auf die weiterzuleitenden Datenwerte angewendet werden und diese entweder verstärken oder abschwächen. Die Berechnungen eines Neuronalen Netzes verlaufen dabei von der Eingabeschicht über die verborgenen Schichten zur Ausgabeschicht. Die Ausgabeschicht

enthält oft nur ein einzelnes Neuron, das eine einfache Antwort mit einem Wert zwischen 0 und 1 liefert (1=Ja, 0=Nein). Mit diesem Ausgabeneuron kann zum Beispiel die Frage „Stellt das Bild einen Laubbaum dar?“ beantwortet werden. Werte zwischen 0 und 1 drücken Unsicherheit in der Antwort aus, 0.9 kann zum Beispiel bedeuten, das Netz ist sich ziemlich sicher, dass es sich um einen Laubbaum handelt, bei einem Wert von 0.1 handelt es sich wohl eher nicht um einen Laubbaum. Besonders bekannt und erfolgreich ist zurzeit das Deep Learning, das gerichtete Netze mit sehr vielen verborgenen Schichten und Knoten verwendet und vor allem auch die Eingabeparameter selbständig aus den Eingabedaten bestimmt. Die Architektur eines Netzes, d.h. die Anzahl der Knoten und Schichten basiert auf empirischen Erfahrungswerten, ebenso wie die geeignete Kodierung eines Problems und ist Gegenstand intensiver Forschungen.

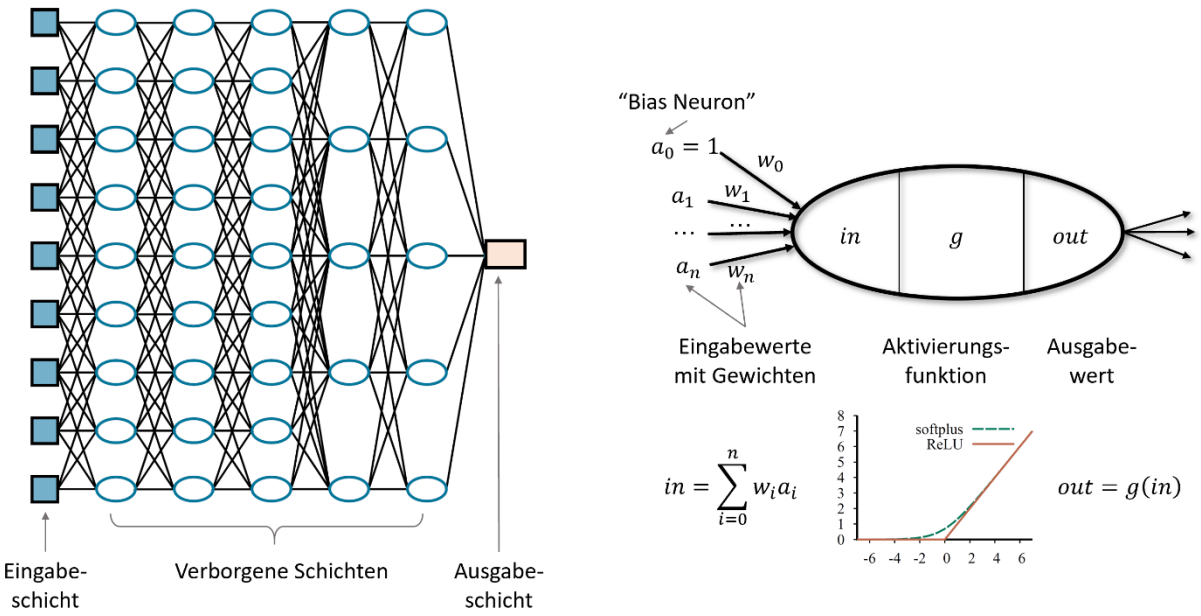

Abbildung 2: Deep Learning: Tiefes Neuronales Netz mit 5 verborgenen Schichten (links) und Berechnungen in einem einzelnen Knoten einer verborgenen Schicht. In Anlehnung an (Russel und Norvig 2020).

Jeder Knoten in einer verborgenen Schicht eines Neuronalen Netzes stellt eine einfache mathematische Berechnungseinheit dar, die auf die Eingabewerte an den eingehenden Kanten eine mathematische Funktion anwendet und den Ergebniswert der Funktion über die ausgehenden Kanten weitergibt. Diese Berechnung illustriert die rechte Seite in Abbildung 2. Als Aktivierungsfunktion können zum Beispiel Schwellwertfunktionen wie die *softplus* oder *ReLU* Funktion verwendet werden: Nur wenn die Summe der eingehenden Daten einen bestimmten Schwellwert überschreitet, wird der Wert der Summe als Ausgabe weitergeleitet, andernfalls werden die Eingaben unterdrückt und es wird 0 weitergegeben.

Man unterscheidet gerichtete Netze, bei denen die Eingabeparameter an den Knoten der Eingabeschicht eingegeben werden, in nur einer Richtung durch alle Zwischenschichten des

Netzes fließen und dann die Ausgabeschicht erreichen. Ungerichtete oder sogenannte Rekurrente (rückgekoppelte) Netze erlauben es, dass Berechnungen in verschiedene Richtungen, also auch von der Ausgabeschicht wieder in die Eingabeschicht, fließen können.

Beim Trainieren eines Neuronalen Netzes werden die Gewichte an den Kanten des Netzes gelernt. Zunächst werden beliebige Gewichte im Netz verteilt. Dann wird ein Trainingsdatensatz in das Netz gegeben und bestimmt, welche Ausgabe das Netz für diesen Trainingsdatensatz berechnet. Weicht diese Ausgabe vom erwarteten Ausgabewert für den Trainingsdatensatz ab, wird die Fehlerabweichung an der Ausgabeschicht des Netzes bestimmt. Anschließend wird dieser Fehler auf der Grundlage einfacher mathematischer Berechnungen über die Knoten der Zwischenschichten verteilt - jeder Knoten hat einen gewissen Anteil am Fehler. Nachdem der Fehleranteil eines Knotens bestimmt ist, können die Gewichte an den eingehenden Kanten zu diesem Knoten angepasst werden. Mit vielen Trainingsdatensätzen lernt das Neuronale Netz so die gewünschte Eingabe-Ausgabe-Funktion indem es die Fehlerabweichung minimiert. Nach Abschluss der Trainingsphase kann dann das Neuronale Netz mit den gelernten Gewichten einfach und schnell zu einem Datensatz den Ausgabewert berechnen. Die Architektur eines Neuronalen Netzes, d. h. die Anzahl der Schichten, der Knoten und ihrer Verbindungen bleibt zurzeit beim Trainieren unverändert und wird vom Menschen auf der Grundlage umfassender Expertise gewählt. Es gibt aber bereits eine umfassende Forschung, die untersucht, wie auch die Netzstruktur gelernt oder angepasst werden könnte, was dann eher der Plastizität des menschlichen Gehirns entspricht, siehe zum Beispiel (Gaier und Ha 2019).

Alle Ansätze des maschinellen Lernens haben das Problem, dass sie nicht garantieren können, wie gut das gelernte System in einer Anwendung funktionieren wird. Eine große Herausforderung stellt dabei die Qualität der Trainingsdaten dar. Diese sollten eine repräsentative Verteilung der Daten in einer Anwendungsdomäne abbilden, aber oft ist diese Verteilung nicht bekannt oder schwer zu bestimmen. Werden Neuronale Netze auf Probleme aus einem Teil des Anwendungsbereichs angewandt, den sie in der Trainingsphase nicht oder nur unzureichend gesehen haben, dann sind fehlerhafte Ausgaben sehr häufig. Dieses Problem wird auch als Concept Drift bezeichnet (Schlimmer und Granger 1986). Da sich die Umwelt eines Systems immer ändern wird, ist ein wiederholtes Trainieren und wenn immer möglich, auch eine Überwachung eines möglichen Concept Drift notwendig. Ebenso können Neuronale Netze sehr leicht manipuliert werden, ohne dass es für den Menschen bemerkbar ist (Fawzi et al. 2017).

Eine praktikable Lösung dieser Herausforderungen und auch des erwähnten Frame Problems ist möglich, wenn KI-Systeme speziell auf eine konkrete Aufgabe zugeschnitten werden und diese Aufgabe in einer klar abgegrenzten Umgebung zu lösen ist. Neuronale Netze haben sich sehr bei der Spracherkennung bewährt und können heute brauchbare Ergebnisse in der Bildverarbeitung oder Sprachübersetzung liefern. Erfolgreich im Einsatz sind Logik-basierte Methoden zum Beispiel im Bereich der Ontologien und Wissensgraphen oder wenn komplexe Probleme auf simple aussagenlogische Formeln mit Millionen von logischen Variablen abgebildet werden. Sogenannte Satisfiability Checker (Biere 2009) prüfen dabei die logische Erfüllbarkeit der resultierenden Formeln und können so zum Beispiel sicherheitskritische Eigenschaften technischer Systeme überprüfen oder komplexe Produkte konfigurieren. Ein typisches Beispiel für fokussierte Anwendungen, bei denen eine Vielzahl unterschiedlicher KI-Methoden inklusive Neuronaler Netze erfolgreich integriert wird, sind Spiele, wie zum Beispiel Schach oder Go, die immer sehr gern von KI-Forschern für die Entwicklung und das Testen

ihrer Algorithmen herangezogen werden. Man spricht in diesem Zusammenhang oft von schwacher KI, also einem System, das dem Menschen vielleicht in einer Spezialaufgabe überlegen ist, d.h. besser Schach spielt, aber bei ganz anderen Aufgaben völlig versagt. Interessant ist dabei, dass sich die Methoden von Mensch und Maschine bei der Lösung dieser Spezialaufgaben nicht nur fundamental unterscheiden, sondern auch sinnvoll ergänzen und gegenseitig befruchten (Bushinsky 2009) - eine Eigenschaft, die für die Anwendung von KI-Methoden in allen wissenschaftlichen Disziplinen spricht und uns vermutlich in der Zukunft weitere fundamentale Erkenntnisse in diesen Disziplinen erlauben wird.

Eine starke KI, heute auch oft als Artificial General Intelligence (AGI) bezeichnet, versteht sich abgrenzend zur schwachen KI eher als eine KI, die dem Menschen in allen Belangen ebenbürtig oder sogar überlegen ist (Goertzel und Pennachin 2007, Goertzel 2014, Mindt und Montemayor 2020). Grundlegende philosophische, aber auch psychologische Fragestellungen spielen innerhalb der AGI damit eine ebenso große Rolle wie konkrete Fragen nach einem simulationsfähigen Modell des menschlichen Gehirns. Interessanterweise kommt hier auch sofort wieder die Frage nach einer Neuinterpretation des Turing Tests auf, da die AGI sich ja im Vergleich zum Menschen misst. So gibt es zum Beispiel den Vorschlag, einen Roboter als Test erfolgreich die Grundschule absolvieren zu lassen. Während die sogenannte schwache KI intelligente Systeme auch ohne Bewusstsein entwickeln kann, stellt sich für die starke KI sofort die Frage nach dem Zusammenhang zwischen (Alltags-)Intelligenz und Bewusstsein, die auch in der Kunst immer wieder thematisiert wird. Steven Spielbergs Film von 2001 bringt die Frage nach einer AGI sehr schön auf den entscheidenden Punkt und lässt seinen Protagonisten im Prolog des Films das entsprechende Forschungsziel formulieren:

„I propose that we build a robot child who can love. A robot child who will genuinely love the parent, or parents it imprints on with a love that will never end. […] You see, what I'm suggesting is that love will be the key by which they acquire a kind of subconscious never before achieved. An inner world of metaphor, intuition, of self-motivated reasoning, of dreams." [Prolog aus Spielberg: Artificial Intelligence, Minute 3:22 - 4:47]

Ein dem Menschen ebenbürtiges KI-System, das in unserer Umwelt erfolgreich agieren kann, bringt uns damit wieder zur Frage, inwieweit KI-Systeme eigene Ziele verfolgen sollen und zwingt uns somit, unsere Haltung zu diesen Maschinen und ihren Rechten zu überdenken. Auch der nachfolgende Beitrag in diesem Buch spricht diese Frage in Form der „Gewöhnungsfähigkeit" an: Die Maschine rückt näher an uns heran. Wie gehen wir damit um?

Spielberg stellt diese entscheidende Frage am Beispiel eines liebenden Roboterkindes. Der Prolog des Films endet mit der Frage einer Kollegin, dargestellt von April Grace, an den Protagonisten:

"You know, it occurs to me with all this animus existing against Mecha's today, it isn't simply a question of creating a robot who can love. But isn't the real conundrum, can you get a human to love them back?" [Prolog aus Spielberg: Artificial Intelligence, Minute 4:52 - 5:07]

Als Antwort erhalten wir hier nur die Aussage des Protagonisten, dass das Roboterkind seinen Menschen immer lieben wird, doch die Kollegin fragt nach:

"But you haven't answered my question. If a robot could genuinely love a person, what responsibility does that person hold toward that Mecha in return?" [Prolog aus Spielberg: Artificial Intelligence, Minute 5:20 - 5:33]

Diese Frage bleibt unbeantwortet. Eindrücklich der Gesichtsausdruck von William Hurt, der eine Mischung von Zweifel, Hilflosigkeit, Verstörtheit und Unverständnis der Frage widerzuspiegeln zu scheint. Der Film geht den Folgen des ungeklärten Verhältnisses zwischen dem Menschen und der intelligenten Maschine nach und auch wir wollen im letzten Teil dieses Beitrags diese Frage anhand heutiger KI-Anwendungen noch etwas weiter beleuchten.

Aktuelle Anwendungen, ethische Fragestellungen und die Zukunft der KI

Sprach- und bildbasierte Technologien haben einen sehr hohen Reifegrad erreicht und wir sehen bereits in einigen Anwendungsgebieten Systeme, die in der Lage sind, Ergebnisse zu erzeugen, die von menschlicher Leistung nicht mehr unterscheidbar sind oder diese sogar übertreffen. Durchbrüche in der Spracherkennung mit Deep Learning haben es ermöglicht, sprachgesteuerte Assistenzsysteme zu entwickeln, denen wir Aufträge erteilen können, z. B. im Navigationssystem des Autos, das übrigens den Weg mit Hilfe von KI-Suchalgorithmen berechnet. Ebenso können wir Fragen zu historischen oder aktuellen Ereignissen oder auch den unterschiedlichsten Wissensgebieten stellen, wie zum Beispiel an eine Alexa auf einem Amazon Echo Gerät. Komplexe Dialoge, die es erfordern auch in einem wechselnden Kontext noch den Sinn einer Frage zu verstehen, können diese Systeme zwar noch nicht führen, aber sie sind doch bereits für verschiedenste Anwendungen sehr gut geeignet. Leider sind sie aber auch weitgehend intransparent für den Benutzer.

Zu einer Maschine sprechen können, verändert die Interaktion zwischen einem Menschen und einer Maschine bereits sehr grundlegend und so stellt sich sofort die Frage, ob eine sprachfähige Maschine für den Menschen immer sofort als Maschine erkennbar sein sollte. Aus ethischer Sicht (High-level expert group 2019) erscheint eine solche Anforderung gerechtfertigt, um die geforderte Transparenz, Fairness und Erklärbarkeit im Umgang mit einem solchen System sicherzustellen. In der aktuellen Realität der verfügbaren Systeme erscheinen diese Eigenschaften nicht vorhanden zu sein, insbesondere der Einsatz vorwiegend weiblicher Stimmen, die in jeder Situation unterwürfig und freundlich bleiben, erscheint problematisch in seiner Auswirkung auf das Frauenbild. Erste Versuche, wie zum Beispiel die genderneutrale Stimme Q stecken noch in den Anfängen und adressieren auch nicht das Ziel, eine maschinentypische Stimme zu entwickeln.

Fussnote: Siehe https://www.genderlessvoice.com

Eine Maschine bei uns zu haben, die uns auf Schritt und Tritt begleitet und viele sinnvolle Dienste anbietet, ist für die meisten Menschen seit dem Durchbruch des Smartphones Realität. Noch trägt der Mensch diese Maschine in Form eines kleinen Kastens herum. Zukünftig werden wir aber Maschinen nutzen, die selbständig mit uns gehen und einen eigenen Aktionsradius und eine gewisse Unabhängigkeit von uns haben. Seit kurzem ist eine solche, bereits recht robuste Maschine auf dem Markt erhältlich. Ob es Anwendungen geben wird, die sich wirtschaftlich rechnen, ist noch offen.

Fussnote: Siehe https://www.bostondynamics.com/spot

Für den weiteren Erfolg solcher Maschinen werden ihre mechanischen Kunstfertigkeiten entscheidend sein. Wie gut können sie greifen oder laufen, d.h. wie gut sind ihre Aktuatoren, sind diese unseren Händen und Beinen ebenbürtig oder überlegen? Wie gut sind ihre Sinnesorgane, d.h. ihre Sensoren, mit denen sie die Umwelt wahrnehmen? Und einerseits zwar

ziemlich banal, aber andererseits marktentscheidend: Wie lange hält die Batterie bzw. wie wird die Energieversorgung sichergestellt? Aber es wird auch darauf ankommen, ob wir diese Maschinen als unterstützend und bereichernd oder als bedrohlich wahrnehmen. Wie werden wir die Grenzen im Spannungsfeld von Nähe und Distanz, Abhängigkeit und Unabhängigkeit definieren?

Aus juristischer Sicht ist es eine äußerst spannende Frage, ob ein Mensch zur Maschine werden kann, wenn er/sie immer weitere Teile des biologischen Körpers ersetzt und ob umgekehrt eine Maschine zum Menschen wird, wenn sie immer mehr menschliche Fähigkeiten erwirbt. Kulminieren tut diese Frage im Problem der Schuldfähigkeit und Verantwortung. Haftet das KI-System für seine Fehler und wie treffen wir bewusste Entscheidungen in kritischen Situationen? Ein autonomes Fahrzeug trifft in einer Situation, die zu einem Unfall führen kann, seine Entscheidungen ganz anders als der Mensch. Während dieser eher unbewusst handelt, sich an seine Überlegungen und Handlungen kaum erinnern kann, entscheidet das autonome Fahrzeug auf der Grundlage deterministischer und stochastischer Berechnungen. Sehr eindrücklich werden die auftretenden Fragen im Moral Machine Experiment des MIT deutlich. In diesem Experiment stimmen Menschen ab, wie ein autonomes Fahrzeug im Falle eines unvermeidbaren Unfalls entscheiden soll. Hier hält uns die Technologie den Spiegel vor und fragt uns nach dem Wert eines Lebens.

Fussnote: Siehe https://moralmachine.mit.edu/

KI-Technologien wie maschinelles Lernen oder Suchalgorithmen ermöglichen neue Anwendungen und Geschäftsmodelle. Ob eine Technologie das Gewünschte leistet und welcher Nutzen und welche Risiken damit verbunden sind, hängt von der Technologie und vom Kontext der Anwendung ab und kann auch nur innerhalb des Anwendungskontexts adressiert werden (Koehler 2018). Es gibt viele Beispiele für den erfolgreichen Einsatz von KI-Technologien in bestimmten Anwendungsbereichen. Es lässt sich aber daraus nicht zuverlässig vorhersagen, dass die gleichen Technologien auch in anderen Bereichen genauso erfolgreich sein werden. Ein Beispiel dafür ist der bereits am Anfang des Artikels erwähnte Erfolg von IBM´s Watson System, das sehr erfolgreich Jeopardy spielen konnte und mit seinen Siegen das aktuelle Interesse an KI ausgelöst hat, aber die Erwartungen an einen Durchbruch in der medizinischen Diagnose und Behandlung nicht erfüllen konnte (Strickland 2019).

Im Bereich der Künste führen KI-basierte Verfahren zu vollkommen neuen Ausdrucksformen, allen voran in der bildenden Kunst. Die von der Kunstkuratorin Marnie Benney 2019 ins Leben gerufene Webseite AIAirtists.org etwa liefert einen Überblick über eine innovative Gemeinschaft von KI-Künstlerinnen und KI-Künstlern, die in ihrer kreativen Praxis auf KI-Technologien setzen und die damit verbundenen gesellschaftlichen und ethischen Herausforderungen in ihren Kunstwerken bereits selbstreflexiv verhandeln.

Fussnote: aiartist.org [zuletzt abgerufen am 20.7.2020)

Neben der Reflektion der eigenen Geschichte (Crevier 1993, Nilsson 2010) setzt sich die KI-Community sehr aktiv mit der Ausrichtung ihrer zukünftigen Forschung auseinander (Gil und Selman 2019, Russell 2018). Dabei ist auch ein Vergleich mit historischen Reflektionen zu den Errungenschaften und Perspektiven der KI durchaus spannend (Hearst und Hirsch 2000,Schurmann 2001,Churchland und Churchland 1990), die sehr viele Ähnlichkeiten mit den aktuellen Diskussionen aufweisen. Neben kritischen Positionen (Brooks 2017, Austin, Rollings und Linden 2017), die einer generellen künstlichen Intelligenz eher skeptisch gegenüber stehen

oder diese zumindest als sehr weit in der Zukunft sehen, finden sich auch Positionen wie die Singularity These von Ray Kurzweil (2005), nach der Computer die menschliche Intelligenz gegen 2045 erreichen und dann Mensch und Maschine zu einer Einheit verschmelzen werden. Bis es soweit ist, warten jedoch noch große Herausforderungen auf die KI-Forschung und seit Jahrzehnten ungelöste Probleme sehr grundlegender Natur müssen geklärt werden, die ich als „die 4 großen A“ zusammenfassen und jeweils an einem Beispiel erklären möchte:

Abstraktion: Menschen können recht einfach Zusammenhänge auf unterschiedlichen Abstraktionsstufen begreifen, beschreiben und erklären. Für KI-Systeme ist ein solcher Wechsel bisher nur sehr wenig verstanden. Das gelernte Wissen in einem Neuronalen Netz ist in der Netzarchitektur bestehend aus bis zu Millionen Knoten und Kanten und den gelernten Kantengewichten repräsentiert - eine Darstellung, die für Menschen unverständlich bleibt und die Erklärbarkeit der Entscheidungen von Neuronalen Netzen bisher nicht ermöglicht. Um die numerische Darstellung des gelernten Wissens in natürlichsprachliche Begriffe abzubilden, müssen wir Abstraktionsprozesse verstehen und berechnen können.

Analogie: Zusammenhänge zwischen zunächst anscheinend nicht zusammenhängenden Dingen zu erkennen, ist eine der Fähigkeiten, die die Kreativität von Menschen ausmacht. Wie sagte Schopenhauer bereits: „Die Aufgabe ist nicht, zu sehen, was noch niemand gesehen hat, sondern zu denken, was noch niemand gedacht hat über das, was alle sehen.“ (Schopenhauer 1851, Seite ??)

Argumentation: Entscheidungen nachvollziehbar zu begründen und anhand von soliden Fakten und Zielen zu rechtfertigen ist eine der Voraussetzungen, um in einem demokratischen Prozess zu Entscheidungen zu kommen, die für die Beteiligten als gut und gerecht empfunden werden können. Sich mit Argumenten auseinandersetzen zu können und diese in das eigene Weltbild zu integrieren, es gegebenenfalls zu revidieren, ist eine der Fähigkeiten, die Intelligenz ausmacht.

Alltagswissen: Erfolgreich den Alltag zu gestalten - von den Auswirkungen der grundlegenden physikalischen Gesetze in unserer Umwelt bis zu den wirtschaftlichen und sozialen Spielregeln einer Gesellschaft - ist eine unabdingbare Voraussetzung für die weitere erfolgreiche Entwicklung der Menschheit. Wollen KI-Systeme hier erfolgreich mitwirken, müssen sie Alltagswissen erwerben, verwenden und auch vergessen können.

Am Ende wird es darauf ankommen, die Frage vom Anfang dieses Artikels aus dem Prolog von Spielbergs Film zu beantworten:

„But what does it amount to?”

[Prolog aus Spielberg: Artificial Intelligence, Minute 2:32 - 2:34]

In der deutschen Fassung des Films heißt es *„Aber was haben wir davon?“* Aus der Sicht der Autorin sollten wir ganz einfach die Künstliche Intelligenz für das *gute gelingende Leben* - auf Englisch *flourishing life* - zum Erfolg bringen (BBC 2019, Karger und Koehler, 2019). Doch wie immer ist es das Einfache, das so besonders schwer zum Gelingen zu bringen ist.

Literaturverzeichnis

Austin, Tom und Mike Rollings und Alexander Linden. „Hype Hurts. Steering Clear of Dangerous AI Myths". Gartner Report G00324274. 03.07.2017.

BBC. "Flourishing in the Age of AI". 2019. http://downloads.bbc.co.uk/mediacentre/flourishing-in-the-age-of-ai.pdf (26. Juni 2020)

Biere, Armin und Marijn Heule und Hans van Maaren und Toby Walsh. *Handbook of Satisfiability*, IOS Press, 2009.

Brooks, Rodney. "Intelligence without representation". *Artificial intelligence* 47.1-3 (1991): 139-159.

Brooks, Rodney. "The Seven Deadly Sins of Predicting the Future of AI". *Technology Review* 120.6, 2017.

Bushinsky, Shay. "Deus Ex Machina- A higher Creative Species in the game of Chess". *AI Magazine* 30.3 (2009): 63-70.

Churchland, Paul M. und Patricia Smith Churchland. „Ist eine Denkende Maschine möglich?" In Künstliche Intelligenz: eine Kontroverse, *Spektrum der Wissenschaft* 3 (1990): 47-54.

Crevier, Daniel. *The Tumultuous History of the Search for Artificial Intelligence*. Harper Collins, 1993.

Epstein, Robert. "The Quest for the Thinking Computer". *AI Magazine* 13.2 (1992): 80-95.

Fawzi, Alhussein und Seyed-Mohsen Moosavi-Dezfooli und Omar Fawzi, und Pascal Frossard. "Universal adversarial perturbations". In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, IEEE 2017: 1765-1773.

Ferrucci, David A. "Introduction to This is Watson." *IBM Journal of Research and Development* 56.3.4 (2012): 1.

Gaier, Adam und David Ha. „Weight Agnostic Neural Networks". *Advances in Neural Information Processing Systems* (2019): 5364-5378.

Gil, Yolanda and Bart Selman. "A 20-Year Community Roadmap for Artificial Intelligence Research in the US". A Computing Community Consortium (CCC) workshop report, 2019.

Goertzel, Ben und Cassio Pennachin (Hrsg). *Artificial General Intelligence*, Springer 2007.

Goertzel, Ben. "Artificial general intelligence: concept, state of the art, and future prospects". *Journal of Artificial General Intelligence* 5.1 (2014): 1-48.

Hayes, Patrick and Kenneth Ford. "Turing test considered harmful". In Proceedings International Joint Conference on Artificial Intelligence (IJCAI). Volume 1 (1995): 972-977.

Hearst, Marty and Haym Hirsch. "AI´s greatest trends and controversies". *IEEE Intelligent Systems*, January/February (2000): 8-17.

High-level expert group on artificial intelligence set up by the European Commission. "Ethical Guidelines for Trustworthy AI". 2019.

Karger, Reinhard und Jana Koehler. „Better L-AI-fe – Künstliche Intelligenz für das Gute Leben". *DFKI-Newsletter* 44 (2019): 6-7.

Koehler, Jana. "Business Process Innovation with Artificial Intelligence: Levering Benefits and Controlling Operational Risks". *European Business & Management*. 4.2 (2018): 55-66.

Kurzweil, Ray. *The singularity is near: When humans transcend biology*. Penguin, 2005.

Legg, Shane und Markus Hutter. „A Collection of Definitions of Intelligence". Technical Report 07-07, IDSIA Lugano, 2007.

McCarthy, John und Marvin L. Minsky und Nathaniel Rochester und Claude E. Shannon: "A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence". August 31, 1955, Nachdruck in *AI Magazine*, 27.4 (2006):12-14.

McCarthy, John und Patrick J. Hayes. "Some Philosophical Problems from the Standpoint of Artificial Intelligence". In *Machine Intelligence 4*. Hrsg. von B. Meltzer und D. Michie. Edinburgh University Press 1969: 463-502.

McCulloch, Warren und Walter Pitts. „A logical calculus of the ideas immanent in nervous activity". *Bulletin of Mathematical Biophysics*, 5 (1943): 115–133.

Mindt, Garett und Carlos Montemayor. A" Roadmap for Artificial General Intelligence: Intelligence, Knowledge, and Consciousness". *Mind and Matter*, *18*.1 (2020): 9-37.

Newell, Allen und Herbert A. Simon. "Computer Science as Empirical Inquiry: Symbols and Search". *Communications of the ACM*, 19.3 (1976):113–126

Nilsson, Nils. *The Quest for Artificial Intelligence – A History of Ideas and Achievements*, Cambridge University Pres, 2010.

Russell, Stuart und Peter Norvig. *Artificial Intelligence – A Modern Approach*, Pearson Education, 2. Auflage (2003), 3. Auflage (2010), 4. Auflage (2020).

Russell, Stuart: Interview about the Long-Term Future of Artificial Intelligence, Artificial Intelligence Podcast by Lex Friedman, https://www.youtube.com/watch?v=KsZI5oXBC0k 2018. (25. Juni 2020)

Schlimmer, Jeffrey C. und Richard H. Granger. „Beyond incremental processing: Tracking concept drift". In *5th National Conference on Artificial Intelligence (AAAI)*, 1986: 502–507.

Schurmann, Kyle. "Artificial Intelligence and Expert Systems - AI is Not Yet Programmed to Love". *How Computers Work, Smart Computing Part 1*, August 2001. http://gec.di.uminho.pt/discip/MaisAC/HCW/si10.htm (25. Juni 2020)

Schopenhauer, Arthur. *Parerga und Paralipomena: kleine philosophische Schriften*. Kapitel 6: Zur Philosophie und Wissenschaft der Natur, Berlin 1851.

Shah, Huma. „Turing's misunderstood imitation game and IBM's Watson success". *Annual Convention of the UK Society for the Study of Artificial Intelligence and Simulation of Behaviour* (AISB), Invited Talk, 2011.

Strickland, Eliza. "How IBM Watson Overpromised and Underdelivered on AI Health Care". *IEEE Spectrum*, April 2019 https://spectrum.ieee.org/biomedical/diagnostics/how-ibm-watson-overpromised-and-underdelivered-on-ai-health-care (25. Juni 2020)

Turing, Alan. “Computing Machinery and Intelligence”. *Mind* 49(1950): 433-460.

Wright, Robert. “Can Machines Think”. *Time Magazine*, 25. März 1996: 50-58.