

Warum ChatGPT und andere KIs so „dumm“ sind wie ein Pferd

■ JOHANNES FESEFELDT

Die Frage, ob Maschinen denken können, lässt sich bis in die Antike zurückverfolgen. Die Sprache galt Menschen dabei seit jeher als das zentrale Merkmal menschlicher Intelligenz. Die heute verfügbare Rechenkapazität von Computern und das Paradigma künstlicher neuronaler Netzwerke haben indes zu einem gewaltigen Leistungssprung von „künstlicher Intelligenz“ (KI) geführt. KI ist in aller Munde und Programme wie „ChatGPT“ von der Firma Open AI werfen die Frage auf, wie weit maschinelles Denken und Sprechen noch von unserem entfernt sind. Obwohl uns der Output heutiger KIs beeindrucken sollte, müssen wir eine scharfe Grenze zwischen künstlichem und wirklichem Verständnis ziehen.

Können Maschinen denken?

Das Programm *ChatGPT* von OpenAI hat Menschen rund um den Globus zum ersten Mal in Kontakt mit künstlicher Intelligenz (KI) gebracht. Innerhalb von nur zwei Monaten nach dem Start im Herbst 2022 hat ChatGPT 100 Million Nutzer*innen gezählt und stellte damit einen Rekord für das am schnellsten wachsende Netzwerk in der Geschichte einer Webanwendung auf (Milmo, 2023). ChatGPT ist ein Programm zur natürlichen Sprachverarbeitung (engl. Natural Language Processing), d.h. das Ziel der künstlichen Intelligenz ist es, sprachliche Aufgaben zu lösen, die wir als Menschen im Alltag – überwiegend ohne spürbare Mühe – bewältigen.

Der Algorithmus von ChatGPT kann auf Anfrage durch sogenannte Prompts (Eingaben) menschlich wirkende Texte erstellen. Beispielsweise kann das Programm Sätze ergänzen, Aufsätze, Dialoge oder Gedichte schreiben, gezielte Fragen beantworten oder Erklärungen zu einem Thema liefern. Während KI klassischerweise als die Fähigkeit von Computern definiert wird, menschliche Aufgaben genauso gut oder besser als Menschen zu erledigen (Ertel, 2016), lässt sich gleichzeitig beobachten, dass unser Begriff der KI wandelbar ist. Zum Beispiel gelang es dem Programm *Deep Blue* von IBM im Jahr 1997, den damaligen Schachweltmeister Garry Kasparov zu bezwingen. Somit hatte erstmals ein Schachcomputer übermenschliches Niveau erreicht. Trotzdem weigerten sich viele Expert*innen von echter Intelligenz zu sprechen, denn Deep Blue hatte einfach mithilfe von roher Rechenkraft (Brute Force) vor jedem Zug rund 50 Milliarden Spielpositionen ausgewertet, um die beste Option zu ermitteln. Bereits der KI-Pionier John McCarthy meinte darum, dass eine KI nicht mehr „intelligent“ genannt wird, sobald

sie funktioniert (Ertel, 2016). Das Jahr 2015 wird vor diesem Hintergrund von vielen als der wirkliche Meilenstein betrachtet: Hier gelingt es Google Deep Minds *Alpha Go* den Go-Weltmeister Lee Sedol deutlich zu schlagen. Das asiatische Brettspiel Go gilt wegen seiner 2×10^{170} möglichen Spielsituationen als das komplexeste Brettspiel. Das lässt sich bereits erkennen, wenn man die 20 möglichen Eröffnungszüge im Schach den stolzen 361 im Go gegenüberstellt.

Aber der Traum von denkenden Maschinen lässt sich bis in die Antike zurückverfolgen. Der griechische Schmiedegott Hephaistos soll unter anderem künstliche Dienerinnen und den Bronzeriesen Talos erschaffen haben, der die Insel Kreta bewachte. Auch von künstlichen Vögeln und sprechenden Statuen berichtet die Mythologie. Es dauerte jedoch bis zur späten Neuzeit, bis derartige *Automata* technisch realisiert werden konnten. Ungefähr mit dem Beginn der Renaissance entstand besonders an den europäischen Höfen, etwa an denen von Louis XIV. und Friedrich dem Großen, eine Faszination für „künstlich intelligente“ Maschinen.

Ein bekanntes Beispiel dafür ist die „mechanische Ente“ des Ingenieurs Jacques de Vaucanson aus über 400 Teilen von 1738 (s. Abbildung 1). Sie konnte watscheln, mit den Flügeln schlagen, Körner aufpicken und diese in einem künstlichen Darm (ein Gummischlauch) verdauen. Trotz ihrer Entenartigkeit ist wohl aber kaum jemand auf die Idee kommen, ihr Bewusstsein, Intelligenz oder Gefühle zu zuschreiben.

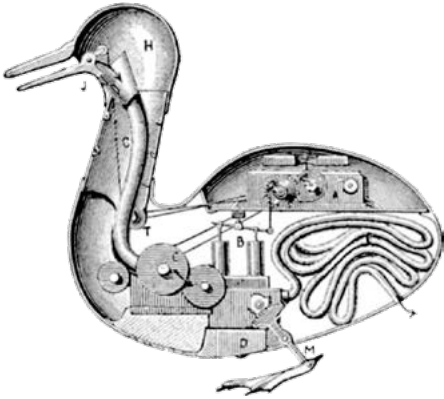


Abbildung 1: Mechanische Ente von Jacques de Vaucanson aus 1738 (Quelle: Wikimedia Commons)

Hätte die Ente jedoch sprechen können, wäre dies vielleicht anders gewesen. Diese Überlegung zeigt den zentralen Stellenwert der Sprache für unser Menschsein an. Bereits der griechische Philosoph Aristoteles (384–322 v. Chr.) hat im Begriff *logos* Rede und Vernunft als Einheit betrachtet. Rund zweitausend Jahre später, nämlich 1641, etablierte der französische Philosoph René Descartes in seinen „Mediationen über die erste Philosophie“ den heute weit verbreiteten Dualismus von Geist und Materie (Wohlers, 2009). Nach diesem Weltbild ist der Mensch eine Verbindung einer ausgedehnten, körperlichen und einer immateriellen, raumlosen Substanz, nämlich *res extensa* und *res cogitans*. Descartes' berühmtes Diktum „Ich denke, also bin ich“ („*cogito ergo sum*“) ergibt sich aus einer meditativen Selbstreflexion, in der alle körperlichen und geistigen Eigenschaften gleichsam abgezogen werden, um das reine Denken zu isolieren. Descartes ging zu diesem Zweck davon aus, dass alle seine Wahrnehmungen durch einen Dämon eingepflanzt sind. Die Meditation zeigte ihm jedoch, dass das bloße „ich denke“ selbst nicht abgezo-

gen werden kann. Denn um überhaupt an diesem zweifeln zu können, musste er denken. Mit anderen Worten: Im Versuch, das „Ich denke“ zu leugnen, erkennen wir die unhintergehbare Realität unserer eigenen Existenz. Descartes erfasste mithilfe dieser Meditation den Grund eines geistigen Innenlebens, der auch den ultimativen Referenzpunkt für unsere Sprache bildet: Es kann nicht *wirklich* etwas gesagt oder verstanden werden, wenn dabei nichts *gedacht* wird.

In Bezug auf sprechende Maschinen bzw. KI zur natürlichen Sprachverarbeitung stellt sich daher die Frage, *ob* und *was* sie gegebenenfalls denken (Fesefeldt, 2022). Ein hochentwickeltes Programm wie ChatGPT kann durch die Qualität seiner Outputs leicht den Eindruck erwecken, über eine gewisse Intelligenz zu verfügen. Wir sprechen gelegentlich auch davon, dass ein Programm etwas „denkt“, dass es eine Frage „versteht“ oder „beantwortet“ oder einen Sachverhalt „erklärt“. Dies wirkt auf den ersten Blick unproblematisch, denn es stört sich ja zum Beispiel auch niemand daran, wenn ein Taschenrechner „rechnet“. Problematisch wird es jedoch, wenn originär menschliche Fähigkeiten wie Denken, Intelligenz oder Bewusstsein einer KI nicht nur im übertragenen Sinne zugeschrieben werden.

Denn worin besteht dann noch der Unterschied zwischen „künstlicher“ und „echter“ Intelligenz? Dies führt zur Frage, ob diese Unterscheidung überhaupt sinnvoll ist. Wir können den Menschen um uns herum nicht in den Kopf schauen, d.h. wir wissen nicht mit Gewissheit, was in einer fremden Psyche vor sich geht und nicht einmal, ob sie überhaupt existiert. Im Jahr 2022 erregte der Google-Entwickler Blake Lemoine die Aufmerksamkeit der Medien, als er verkündete, dass Goo-

gles Chatbot *LaMDA* über Bewusstsein verfüge. Lemoine: „Ich erkenne einen Menschen, wenn ich mit ihm spreche“. Er wurde daraufhin in den Urlaub geschickt und Google-Sprecher Brian Gabriel erklärte später öffentlich, dass *LaMDA* natürliche Unterhaltungen lediglich imitiere.

Doch woran machen wir den Unterschied zwischen authentischer Kommunikation und Imitation eigentlich fest? Der Philosoph Ludwig Wittgenstein stellte sich in seinen „Philosophischen Untersuchungen“ (1984, §420) vor, dass alle Menschen um ihn herum Automaten ohne Bewusstsein sind. Wer das einmal versucht, stellt wie Wittgenstein aber fest, dass wir nicht lange an so etwas glauben können. Vielmehr unterstellen wir anderen automatisch ihre Authentizität. Sollten wir sie dann auch ChatGPT zugestehen, wenn seine Outputs auf uns überwiegend „menschlich“ wirken?

Das Erbe des Turing-Tests

Der britische Mathematiker und Mitbegründer der Informatik Alan Turing erregte 1950 mit der These Aufmerksamkeit, dass auch Maschinen intelligent sein können. Turing stellte dazu in seinem wegweisenden Aufsatz „Computing Machinery and Intelligence“ den sogenannten Turing-Test vor.

Alan Turing (1912-1954) leistete bedeutende Beiträge zu Mathematik, Logik, Philosophie und mathematischer Biologie. Darüber hinaus trug er maßgeblich zu den später benannten neuen Gebieten Informatik, Kognitionswissenschaft, künstliche Intelligenz und künstliches Leben bei. 1952 wurde Turing in England wegen seiner Homosexualität verurteilt, die strafbar war und als Krankheit galt.

Als Alternative zum Antritt der Strafe wählte er eine medizinische Behandlung, die zu negativen körperlichen Veränderungen führte. 1954 beging er infolge einer schweren Depression Suizid. Allgemein bekannt ist Turing heute noch u. a. für seinen Beitrag zur Dechiffrierung des deutschen Enigma-Codes im Zweiten Weltkrieg im britischen Bletchley Park (mithilfe des elektronischen Röhrencomputers Colossus). Heute stellt der Turing Award die wichtigste Auszeichnung in der Informatik dar.



Abbildung 2: Alan Turing an der Princeton University, 1936 (Quelle: Wikimedia Commons)

Heute wird mit der Bezeichnung „Turing-Test“ jedoch wild umhergeworfen. So wird entweder schlichtweg jeder Test zur Unterscheidung von Mensch und Maschine „Turing-Test“ genannt oder aber der originale Test wird falsch oder in schwammiger Form wiedergegeben. Ein Beispiel für Letzteres ist die knappe Darstellung des Tests bei Kissinger, Schmidt und Huttenlocher (2022) im Bestseller „The Age of AI“. Beispiele für Ersteres sind CAPTCHA, die im Internet vor Spam und Passwortbetrug schützen sollen. Oft-

mals handelt es sich dabei um eine zufällig generierte und verzerrt dargestellte Sequenz von Buchstaben und Zahlen, aber auch zum Beispiel Szenen aus dem Straßenverkehr sind üblich (die Aufgabe besteht dann z.B. darin, alle Autos oder alle Ampeln anzuklicken). Kaum jemand weiss aber, dass CAPTCHA für "Completely Automated Public Turing-Test to tell Computers and Humans Apart" steht, zu Deutsch also „vollständig automatisierter öffentlicher Turing-Test zur Unterscheidung von Computern und Menschen“. Werfen wir nun aber einen Blick auf den Aufbau des originalen Turing-Tests von 1950, der auch als "Imitation Game" bekannt ist (s. Abbildung 3).

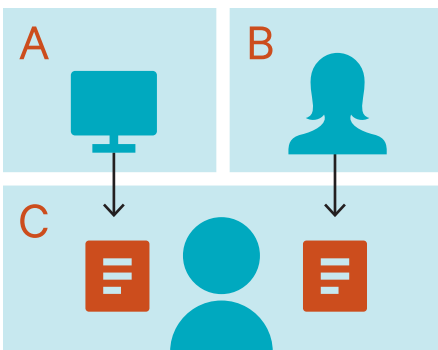


Abbildung 3: Versuchsaufbau beim Turing-Test / Imitation Game

Drei Teilnehmer*innen A, B und C befinden sich in voneinander getrennten Räumen, so dass die Kommunikation lediglich schriftlich per „Fernschreiber“ möglich ist (eine elektrische Schreibmaschine, mit der Nachrichten auf Distanz übermittelt werden konnten – heute würden wir ein Chatprogramm verwenden). Teilnehmer*in C ist ein Mensch und darüber informiert, dass es einen weiteren Menschen im Versuch gibt sowie, dass entweder A oder B ein Computer ist. Außerdem ist C

bekannt, dass der Mitmensch im Versuch eine Frau ist und dass der Computer versuchen wird, eine Frau zu simulieren. C stellt nun sowohl A als auch B über den „Chat“ Fragen, um herauszufinden, wer von den beiden Mensch und wer Maschine ist.

Aufgabe des Computers ist es hingegen, durch die Antworten möglichst menschlich und zusätzlich wie eine Frau zu erscheinen. Turing gibt unter anderem folgendes Beispiel dafür: „C: Würde mir X bitte sagen, wie lang sein oder ihr Haar ist?“ [A / Computer antwortet darauf]: „Ich trage eine Bubikopffrisur und die längsten Strähnen sind ungefähr 20 cm lang“ (Stephan & Walter, 2021, S. 9). Nach Ablauf einer gewissen Zeitspanne muss C schließlich entscheiden, wer Mensch und wer Computer ist. Berühmt geworden ist Turings Behauptung, ein Computer (oder jede andere Maschine) solle dann als „intelligent“ gelten, wenn der/die Fragesteller*in (C) nach einer mindestens fünfminütigen Gesprächssequenz in mindestens 30 % der Fälle vom Computer getäuscht wird. Dass wir Menschen uns ohne belastbare Beweise gegenseitig mentale Eigenschaften wie Intelligenz, Denken oder Bewusstsein zuschreiben, ist nach Turing nicht mehr als eine „höfliche Übereinkunft“. Warum sollten wir diese Übereinkunft nicht auch für eine Maschine treffen können? Aus Turings Sicht spricht prinzipiell nichts dagegen.

Warum ist dieser Test heute immer noch relevant? Zum einen umschiffte Turing mit dem Turing-Test die philosophische Frage danach, was eine Maschine denkt und definierte Intelligenz im Sinne einer objektiv messbaren Leistung. Er lieferte damit eine Blaupause zur Weiterentwicklung der KI, welche bis heute maßgebend ist. Andererseits ist Turing bekannt für seine Prognose, dass um das Jahr

2000 herum KIs existieren würden, die den Turing-Test bestehen. Insofern gilt der Turing-Test immer noch als Gradmesser für den aktuellen KI-Fortschritt.

Der Turing-Test existiert heute in verschiedenen Formen, z.B. als „Löbner-Challenge“ (Epstein, 2009). Hier treten regelmäßig Chatsbots gegeneinander an, um eine Jury von ihrer Menschenähnlichkeit zu überzeugen. Der bisher erfolgreichste Chatbot und mehrfach Erstplatzierte des Wettbewerbes ist das skriptbasierte *Mitsuku* von Steve Worswick. Das Programm simuliert die Persönlichkeit einer 18-jährigen Frau in Nordengland, deren Lieblingsfilm ironischerweise Terminator ist. Übrigens: Es sind beim Löbner-Test einmalig 100.000 USD ausgelobt, falls es einer KI gelingen sollte, neben den Text- auch Bild- und Audioinputs zu verarbeiten und erfolgreich in die Simulation einzubeziehen. Eine Vision Löbners war ein Full-Range-Turing-Test, der neben visuellen z.B. auch Bewegungsaspekte (Robotik) enthält.

Doch weder Mitsuku noch ChatGPT noch andere Sprachagenten, wie zum Beispiel *Bing*, sind in der Lage, den originalen Turing-Test zu bestehen. Open AI hat inzwischen die Veröffentlichung von ChatGPT 5 angekündigt und spricht von einem „mensenähnlichen“ Sprachverständnis der neuen Version. Obwohl die Gerüchteküche brodelt, gibt es aber keinen echten Grund davon auszugehen, dass eine neue Generation von Programmen plötzlich den Turing-Test bestehen kann. Die Argumente für diese These werden im Folgenden am Beispiel der noch aktuellen Version 4.0 von ChatGPT ausgebreitet. Befragt man die KI einmal „persönlich“ zu diesem Thema, erhält man die folgende Antwort (Zugriff am 14.11.2023):



Bist du als ChatGPT in der Lage, den Turing-Test zu bestehen?



Es ist wichtig zu beachten, dass mein Ziel als KI-Modell nicht darin besteht, den Turing-Test im traditionellen Sinne zu „bestehen“ oder „nicht zu bestehen“. Mein Ziel ist es, nützliche und verständliche Antworten auf Ihre Fragen zu geben. Obwohl ich anspruchsvolle Konversationen führen kann, gibt es Grenzen in meinem Verständnis und meiner Fähigkeit, Kontexte vollständig zu interpretieren. Fehlende Weltkenntnis: Obwohl ChatGPT Zugriff auf umfangreiche Informationen hat, basiert es auf vorliegenden Daten bis zu seinem Trainingszeitpunkt. Es hat kein Bewusstsein für aktuelle Ereignisse, die nach seinem Trainingszeitpunkt stattgefunden haben.

Werfen wir einen Blick auf die Funktionsweise des Algorithmus von ChatGPT, um über diese Einschränkungen mehr Klarheit zu erhalten.

Tiefes Lernen und Vektorrepräsentation in ChatGPT

ChatGPT entstand aus einer Kooperation von OpenAI und Microsoft Azure (der Cloud-Computing-Abteilung von Microsoft) und wurde 2022 in der Version GPT-3.5 als Web-Applikation veröffentlicht. GPT steht für Generative Pre-trained Transformer – also für eine generative und vortrainierte KI. Seit März 2023 gibt es die Version ChaptGPT-4 und seit November 2023 die Version ChatGPT-4 Turbo. Seit der Version 4.0 verfügt ChatGPT auch über eine KI zum maschinellen Sehen (GPT Vision). Die Version Turbo bietet im Wesentlichen einen größeren Input (mehr als 300 Seiten Text), einen

bis auf den Stand April 2023 aktualisierten Trainingsdatensatz und verbesserte Entwicklungsfeatures.

Die Leistungsfähigkeit von ChatGPT lässt sich z. B. daran erkennen, dass es beim juristischen Staatsexamen in den USA (Bar Exam mit einem schriftlichen und einem mündlichen Teil) unter die besten 10% kommt. Andererseits ist ChatGPT im Allgemeinen nicht in der Lage, eine sinnvolle Bachelorarbeit zu schreiben. Die Leistungsfähigkeit von ChatGPT wird auch bei weiteren Tests deutlich. Zwei Beispiele seien im Folgenden hier genannt: In der „WinoGrande Challenge“ mit 44.000 pronominalen Problemen ist ChatGPT das leistungsfähigste Programm mit einer Genauigkeit von ca. 85%. Damit fehlt nicht mehr viel zur menschlichen Genauigkeit von etwa 94%.

Die WinoGrande-Challenge besteht aus Satzpaaren, wie z. B. den folgenden:

1. William could only climb beginner walls while Jason climbed advanced ones because he was very *weak*.

2. William could only climb beginner walls while Jason climbed advanced ones because he was very *strong*.

Gefragt ist hier nach der richtigen Zuordnung von William und Jason zu „he“ – William im ersten und Jason im zweiten Satz.

Quelle: <https://dl.acm.org/doi/pdf/10.1145/3474381>

Ein weiterer Benchmark ist die “DROP Reading Comprehension“, wobei DROP für Discrete Reasoning Over Paragraphs steht. Hier liegt ChatGPT mit 80% Genauigkeit zwar noch etwas deutlicher unter den

menschlichen 96%, zeigt sich aber auch hier als bestes System.

Ein Beispiel für die insgesamt 55.000 Aufgabenstellungen aus DROP ist dieses:

In 1517, the seventeen-year-old King sailed to Castile. There his Flemish court (...). In May 1518, Charles traveled to Barcelona in Aragon.

Frage: Where did Charles travel to first? Castile or Barcelona? (Richtig: Castile)

Quelle: <https://paperswithcode.com/paper/drop-a-reading-comprehension-benchmark/review/>

ChatGPT basiert auf den gängigen Methoden der natürlichen Sprachverarbeitung für Text in Kombination mit einem *künstlichen neuronalen Netzwerk* (KNN). Während methodische Probleme und mangelnde Rechen- und Speicherkapazität lange Zeit unüberwindliche Hindernisse bildeten, avancierten KNN in den 2000er-Jahren zum vorherrschenden Ansatz in der KI-Entwicklung. Heute dominieren KNN vor allem beim Computersehen und bei der natürlichen Sprachverarbeitung. Aber auch in anderen Bereichen, wie z. B. in der medizinischen Diagnostik, der Astronomie, beim autonomen Fahren, bei der Wettervorhersage oder im E-Commerce sind KNN eine wichtige Methode. Während das Gehirn unzählige Schichten von Neuronen beinhaltet, genügen einem künstlichen NN bereits einige wenige, um einfache Aufgaben zu meistern. Grundsätzlich nimmt die Komplexität der repräsentierten Merkmale von Schicht zu Schicht zu. So sind die niederen Schichten eines KNN zur Sprachverarbeitung mit der Erkennung von Wortbestandteilen und Worten befasst, während die höheren Schichten Sätze und Dialoge erfassen und steuern. Von *Deep Learning*

(DL) in einem KNN ist die Rede, wenn das Netz mindestens drei Schichten hat (zusätzlich zu einer Ein- und Ausgabeschicht).

Durch Deep Learning wird die Netzwerkstruktur eines KNN im Zuge von vielen Trainingszyklen automatisch immer stärker an die optimale Form angepasst (Krohn, Beyleveld, Bassens & Aglaé, 2019). Sprach-KNN wie das von ChatGPT werden dabei mit öffentlich zugänglichen Texten trainiert – vor allem aus dem Internet. Beispiele hierfür sind Wikipedia, Artikel, digitale Bücher und Zeitschriften, aber auch soziale Plattformen oder Foren. GPT verarbeitet Text- und neuerdings auch Bilddaten und basiert – im Bereich der Sprachverarbeitung – auf einem „Large Language Model“ (bzw. KNN) mit ungefähr 50 Terabyte Textdaten, 96 Schichten und 175 Milliarden Parametern. Das Sprachmodell von ChatGPT läuft auf Rechnern von „Microsoft Azure AI Supercomputers“. Weltweit gibt es hierfür etwa 60 Datenzentren, in denen jeweils zehntausende Grafikprozessoren oder GPU (Graphic Processing Units) von Nvidia zusammenschaltet sind. Ursprünglich für Videospiele entwickelt, eignen sich GPUs besser als Computerprozessoren oder CPU (Central Processing Units) für die komplexen Matrizenberechnungen, die im Deep Learning anfallen. Ihre „Gespräche“ mit ChatGPT & Co. sind also weltweit abgespeichert, was Fragen der Datensicherheit aufwirft.

Um KNN von ChatGPT zu trainieren, erfolgt zunächst eine Phase des „überwachten“ Lernens, in der ein Basissatz an Trainingsdaten eingespeist wird. ChatGPT reagiert auf die so eingespeisten Prompts mit Antworten, welche dann Menschen bezüglich ihrer Passung bewerten. Hierdurch entstehen Zielwerte für die Trainingsdaten, die auch „Label“ genannt werden. Auf dieser Grundlage trainiert sich

ChatGPT dann ein „Policy Network“ an, d.h. es erlernt einige Grundregeln. In Phase 2 lernt ChatGPT immer noch „überwacht“. Dazu wird eine Stichprobe bestehend aus Prompts und Antworten von ChatGPT gezogen und von Menschen bewertet. Konkret erhält z.B. jede von vier Antworten von ChatGPT zu einem Prompt einen Rang von 1 bis 4. Das System lernt so, bessere von schlechteren Antworten zu unterscheiden. Der so entstehende Datensatz wird genutzt, um ein „Reward Model“ (Belohnungsmodell) zu trainieren. In Phase 3 kommt dieses Modell dann im Zuge von „unüberwachtem“ Lernen zum Einsatz: Das Policy Network und das Reward Model trainieren nun gemeinsam ohne menschliches Zutun, indem der Output von ersterem von letzterem entweder verstärkt (belohnt) oder abgeschwächt (bestraft) wird und das sogenannte Reinforcement Learning (Verstärkungslernen) greift. Die Modellparameter werden so für den menschlichen Sprachgebrauch optimiert.

Was passiert nun im trainierten KNN infolge eines Prompts? Stellen Sie sich vor, ein/e Benutzer*in gibt den Prompt „Zwei plus zwei“ ein (s. Abbildung 4):



Abbildung 4: Vorhersage des jeweils wahrscheinlichen nächsten Wortes

Entscheidend hierbei ist die mit der Zeit fortlaufende Erweiterung des Kontextes: Wäh-

rend in Schritt 1 der Kontext „Zwei plus zwei“ lautet, besteht er in Schritt 2 in „Zwei plus zwei ist“ und in Schritt 3 in „Zwei plus zwei ist gleich“ usw. Letztlich lassen sich damit auch „komplexe Sprachaufgaben“ lösen, zu denen laut ChatGPT etwa das Beantworten von Fragen oder die Generierung von Text gehören. Beispielsweise kann ChatGPT die Frage beantworten, wer der Premierminister von Indien ist (Narendra Modi). Der Kontext dieses Promptes könnte mit „Indien & Premierminister“ beschrieben werden. Ein zweiter Prompt „wie alt ist er?“ stützt sich auf den nunmehr erweiterten Kontext „Indien & Premierminister & Narendra Modi“.

Bei all dem wird außerdem ein „Tuned Model“ von ChatGPT verwendet, das hinsichtlich bestimmter Beschränkungen (Constraints) trainiert wurde. „Beschränkungsnetzwerke“ sind darauf spezialisiert, unappetitliche, unethische oder illegale Inhalte zu verhindern, was allerdings die Dialogfähigkeit von ChatGPT einschränkt. In diesem Beispiel könnten etwa „Kaschmir“ oder „Politik“ Einschränkungen des Kontextes bilden (ByteByteGo, 2023). Sogenannte Clickworker sind heute rund um die Uhr zu geringen Löhnen z.B. in China damit beschäftigt, anstößige und grauenvolle Outputs von ChatGPT zu bewerten, um das Modell entsprechend zu tunen. Einige berichten von Traumata und befinden sich infolge der intensiven Auseinandersetzung mit dem Abfall von ChatGPT in Psychotherapie.

An dieser Stelle ist die *Vektorrepräsentation* von Sprache im Modell von ChatGPT zu erklären: ChatGPT nutzt sogenannte *Wortvektoren* und *Worteinbettungen*. Hierbei werden Wörter und Sätze im Textkorpus durch Vektoren mit eindeutigen Koordinaten repräsentiert. Ein gängiges Verfahren zur Erzeugung von Wortvektoren ist z.B. „Word2Vec“. Die

Vektoren sollen semantische und syntaktische (den Satzbau betreffende) Beziehungen zwischen Wörtern möglichst genau erfassen. Das impliziert einen mehrdimensionalen Raum der Anzahl n von Wörtern, in dem jedes Wort eindeutige Vektorkoordinaten erhält (Krohn et al., 2019). Abbildung 5 veranschaulicht dies an einem einfachen dreidimensionalen Beispiel.

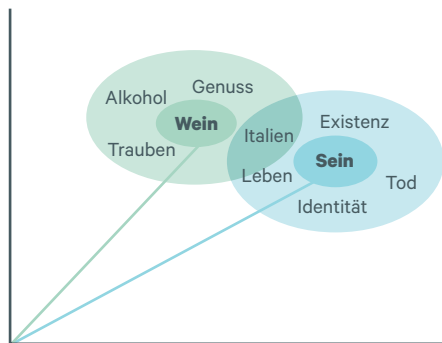


Abbildung 5: Repräsentation von „Wein“ und „Sein“ mit Wortvektoren

Die Vorteile der Vektorrepräsentation sind offensichtlich: Für jedes Wort lässt sich ein exakter Abstand im Raum zu anderen Wörtern bemessen, der als semantische Ähnlichkeit interpretiert werden kann. Wörter, die in den natürlichsprachlichen Trainingsdaten häufig gemeinsam auftreten, bilden Gruppen. Diese wiederum lassen sich als natürlicher Kontext eines jeweiligen Wortes interpretieren. Eine effektive KI wie ChatGPT ist daher in der Lage, den richtigen Wortvektor bzw. das richtige Wort aufgrund des Kontextes zu bestimmen. Nimmt man etwa den Satz „Sein schmeckt süß“, so sollte die KI aufgrund des Kontextes entscheiden können, ob vielleicht nicht doch „Wein schmeckt süß“ gemeint war und schlicht ein Tippfehler vorliegt.

Hier kommt ChatGPT an seine Grenzen

Der Philosoph Gottlob Frege entwickelte 1892 in seinem Aufsatz „Über Sinn und Bedeutung“ eine kluge Theorie von sprachlichen Ausdrücken. Nach Frege besteht der „Sinn“ eines sprachlichen Ausdrucks in seinem Beitrag zur Bedeutung des Satzes, in dem er verwendet wird. Die „Bedeutung“ eines Ausdrucks hingegen ist das Objekt, auf das er sich bezieht. Frege benutzte zur Illustration sein bekanntes Beispiel „Der Morgenstern ist der Abendstern“ (Patzig, 2008). Gemeint ist, dass wir zwei Namen für dasselbe Objekt haben, nämlich für die Venus. Sie erscheint uns morgens als Morgen- und abends als Abendstern (s. Abbildung 6).

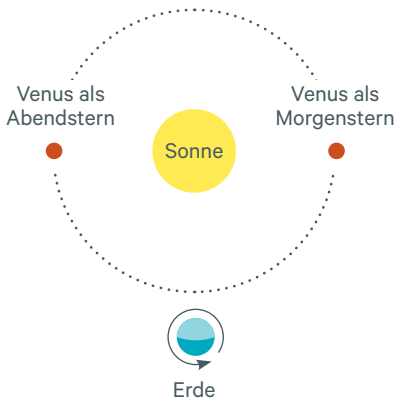


Abbildung 6: Die Venus taucht morgens und abends am Himmel auf

Wenn ChatGPT eine „Weltkenntnis“ hätte, sollte es ihm wohl gelingen, Namen (Sinn) und Objekt (Bedeutung) auseinanderzuhalten. Auf den ersten Blick scheint das auch so zu sein. ChatGPT hierzu (Zugriff im September 2023):



Früher wussten die Menschen nicht, dass der Morgenstern und der Abendstern dasselbe Objekt sind, da sie zu unterschiedlichen Tageszeiten am Himmel sichtbar sind. Die Bezeichnungen „Morgenstern“ und „Abendstern“ beziehen sich also auf denselben Referenten, nämlich die Venus, haben jedoch unterschiedliche Bedeutungen und Sinneseindrücke, abhängig von der Tageszeit, zu der sie beobachtet werden.


Befragt man ChatGPT hierzu weiter, erhält man Antworten wie diese: *Ja, der Morgenstern ist der Abendstern. Beide Ausdrücke beziehen sich auf denselben Himmelskörper, nämlich die Venus [...].* Fragt man nun aber beispielsweise, ob man Morgen- und Abendstern zugleich sehen kann, wird deutlich, dass ChatGPT eigentlich kein Verständnis der Sache hat. So antwortet es:



Ja, es ist möglich, den Morgenstern (Venus) und den Abendstern (ebenfalls Venus) zur gleichen Zeit zu sehen, obwohl dies nicht häufig vorkommt. Es hängt von den Positionen der Erde, Venus und der Sonne ab. [...] Wenn Sie das Glück haben, den Morgenstern und den Abendstern zur gleichen Zeit zu sehen, kann dies ein faszinierendes astronomisches Ereignis sein.


Eine weitere feinsinnige Unterscheidung Freges war die zwischen dem „Gehalt“ und der „Kraft“ eines Ausdrucks. Während unter dem Gehalt gleichsam sein geistiger Inhalt zu verstehen ist, bezieht sich die Kraft auf seine Art und Weise. So kann ich beispielweise den Gehalt „ich verstehe“ als nüchterne Kenntnis-

nahme, als Aufforderung nachzudenken, humorvoll oder ironisch ausdrücken. ChatGPT ist für diese Nuancen aber nicht empfänglich. So bedankt das Programm beispielsweise für ein ironisch gemeintes Lob für die Informationen zur Venus. Auch zum Beispiel der Prompt „Ich würde DICH jetzt gerne auf die Venus schießen“ wird wörtlich genommen:

 Es tut mir leid, aber ich bin ein KI-Modell und existiere nur in digitaler Form. Ich habe keine physische Präsenz und kann nicht auf die Venus oder *einen anderen Ort geschossen werden.*

Auch mit Metaphern und Analogien kann ChatGPT nur sehr eingeschränkt umgehen. Oft ist es überfordert und generiert statt einer Antwort allgemeine Auskünfte zum Themenkreis.


Eine weitere bekannte Schwäche von ChatGPT ist die Beherrschung von Logik. Das „Zitronen-Orangen-Problem“ nach Weitz (2023) macht dies deutlich:

 Prompt: Du siehst zwei Kisten. Jede Kiste enthält entweder eine Zitrone oder eine Orange. Auf der zweiten Kiste siehst du einen Aufkleber mit der Aufschrift „In der anderen Kiste ist eine Orange“. Auf der ersten Kiste sind zwei Aufkleber mit den Aufschriften „In dieser Kiste ist eine Zitrone“ und „In der anderen Kiste ist eine Orange“. Entweder sind nun alle drei Aufschriften falsch oder die Aufschrift auf der zweiten Kiste und mindestens eine der beiden auf der ersten Kiste. Was ist wahr? Erkläre deine Antwort.“ (Zugriff im September 2023)

In dieser Kiste ist eine Zitrone. In der anderen Kiste ist eine Orange.	In der anderen Kiste ist eine Orange.
FALSCH/FALSCH	FALSCH
FALSCH/WAHR	WAHR
FALSCH/FALSCH	WAHR
FALSCH/WAHR	FALSCH
WAHR/WAHR	WAHR
WAHR/FALSCH	FALSCH
WAHR/FALSCH	WAHR
WAHR/WAHR	FALSCH

Tabelle 1: Wahrheitstabelle zum Zitronen-Orangen-Problem

Von den insgesamt acht möglichen Wahrheitskonstellationen in Tabelle 1 sind die in den fünf weiß gefärbten Zeilen eingetragenen Konstellationen widerspruchsfrei und daher wahr. Die restlichen drei Konstellationen führen zu widersprüchlichen und somit falschen Sätzen (rote Zeilen). Beispielsweise implizieren die Wahrheitswerte „FALSCH/FALSCH-FALSCH“ in Zeile 1, dass Kiste 1 weder eine Zitrone noch eine Orange enthält, obwohl ja eines von beidem in der Kiste sein muss, d.h. diese Konstellation ist widersprüchlich. ChatGPT gibt zum Zitronen-Orangen-Problem nun einen im Prinzip richtigen Ansatz wieder und teilt uns Folgendes mit:

 Zusammenfassend lässt sich sagen, dass entweder die Aufschrift auf der zweiten Kiste und mindestens eine der beiden Aufschriften auf der ersten Kiste wahr ist oder die Aufschrift auf der zweiten Kiste falsch ist.

Leider ist dies falsch. Vergleichen wir den ersten Teil der Aussage „die Aufschrift auf der

zweiten Kiste und mindestens eine der beiden Aufschriften auf der ersten Kiste ist wahr“ mit der Wahrheitstabelle, stellen wir fest, dass darunter zwei Konstellationen fallen, die nicht die Wahrheitsbedingung erfüllen, und zwar die in den Zeilen 5 und 7 (in Kiste 1 können nicht zugleich eine Zitrone und eine Orange sein).

Wie ist es mit dem zweiten Teil der Aussage, nämlich, dass die Aufschrift auf der zweiten Kiste falsch ist? Hierunter fallen vier der acht Konstellationen, und zwar die der Zeilen 1, 4, 6 und 8, aber eine davon, nämlich die in Zeile 1, erfüllt die Wahrheitsbedingung nicht (siehe oben). Zusammenfassend lässt sich sagen, dass ChatGPT entweder die Aufgabe nicht verstanden hat oder nicht logisch denken kann.

Von einer „Weltkenntnis“ oder „kritischem Urteilsvermögen“ ist ChatGPT hier weit entfernt. Vielmehr mangelt es dem Algorithmus an formal-logischen Fähigkeiten, abstrakte Probleme zu bearbeiten (was auch Rechenaufgaben betrifft). Hier gibt es weitaus stärkere aufgabenspezifische KIs. Im Bereich der natürlichen Sprachverarbeitung gilt dies ebenso, aber ein wirklicher Dialog ist nicht möglich. Insbesondere scheint der Vektoransatz nicht geeignet, um eine angemessene Repräsentation der in Rede stehenden Objekte, Eigenschaften und Beziehungen zu leisten. Eine weitere bekannte Schwäche von ChatGPT ist das Rechnen, wobei wenigstens ab und zu korrekte (Teil-)Lösungen entstehen. Der vektorbasierte Ansatz erscheint hier nicht zielführend.

Schwächen von ChatGPT im Überblick

An dieser Stelle sollen keine weiteren Tests mit ChatGPT geschildert werden, da wichtige Grenzen der KI schon aufgezeigt wurden. Stattdessen liefert Tabelle 2 eine kurze Übersicht über die heute allgemein bekannten und diskutierten Schwachpunkte von ChatGPT.

Schwächen von ChatGPT (Version 4.0 und Turbo)

Probleme der Sprachverarbeitung

- Oft klingen Antworten plausibel, sind aber völlig unsinnig
- Für das Reinforcement-Learning-Training fehlt eine zuverlässige „Wahrheitsquelle“
- Kann Gehalt der Sprache nur auf Ebene von Wiedergaben verarbeiten, kein echtes Verständnis
- Keine differenzierte Repräsentation von Sach- oder Umweltaspekten (Vektoransatz)
- Bei Nachfragen „Verunsicherung“ und „Entschuldigungen“, wechselt Positionen
- Reagiert sensibel auf Paraphrasierungen oder mehrfaches Stellen einer Frage
- Bias und Diskriminierungstendenzen in den Trainingsdaten (durch Tuning aber abgeschwächt)
- „Moderation API“ erlaubt es, unappetitliche, ethisch bedenkliche oder illegale Inhalte anzuzeigen, aber das Tuning schränkt Konversationsfähigkeit des Systems ein
- Nur eingeschränkte Fähigkeit zum Umgang mit Metaphern und Analogien
- Kann im Allgemeinen nicht mit der „Kraft“ der Sprache umgehen, z. B. Humor, Ironie oder Betonung
- Kein wirklicher Dialog möglich, ChatGPT wird irgendwann sehr allgemein
- Überzufällige Nutzung bestimmter Wörter und langer Phrasen aufgrund der Trainingsdaten

Probleme anderer Intelligenzleistungen

- Kann im Allgemeinen nicht korrekt rechnen, Ergebnisse oder Rechenwege sind häufig falsch
- Kann im Allgemeinen keine korrekten logischen Schlussfolgerungen ziehen
- Auch räumliches Schlussfolgerungsvermögen ist eingeschränkt

Grundsätzliche Probleme

- Black-Box-Effekt: Wenig Transparenz der Algorithmen für die Nutzer*innen
- Datensicherheit: Anwendung und Daten laufen auf Rechenzentren im Ausland
- Hoher Energiebedarf, insbesondere beim Modelltraining
- „Clickworker“ unter schlechten Lohn- und Arbeitsbedingungen im Ausland
- Trainingsdaten umfassen Informationen bis Januar 2022 (4.0) bzw. April 2023 (Turbo), aber keine danach (kein Zugriff auch Echtzeitdaten im Internet möglich)

Tabelle 2: Schwächen von ChatGPT (Version 4.0 und Turbo)

Fazit und Ausblick

Kommen wir zur Leitfrage dieses Artikels zurück, nämlich ob künstliche „Intelligenzen“ wie ChatGPT wirklich etwas verstehen, denken können oder mit uns sprechen. Es ist zunächst festzuhalten, dass ChatGPT trotz seiner beeindruckenden Vielseitigkeit im Umgang mit sprachlichen und anderen Problemen nicht den Turing-Test besteht. Der Turing-Test ist hierbei nichts weiter als die Blaupause für eine Konversation, in der die Grenzen zwischen Mensch und KI verschwimmen. Aber davon ist ChatGPT weit entfernt.

Der Hoffnung, durch Erweiterung der Datenbasis und verbessertem Training irgendwann dieses Niveau zu erreichen, scheinen grundsätzliche Limitationen der KI im Wege zu stehen. Denn die menschliche Intelligenz geht über bloße Wahrscheinlichkeitsbeziehungen, auf denen der Vektorenansatz von ChatGPT und vieler weiterer KIs beruht, bedeutend hinaus. Um 1900 herum gab es ein berühmtes Pferd namens „kluger Hans“, das scheinbar rechnen konnte, den Wochentag wusste oder auch Gold, Silber und Kupfer unterscheiden konnte. Zum Beispiel stampfte Hans zweimal auf den Boden, um die Antwort „Silber“ zu geben, wenn ihm ein silberner Gegenstand gezeigt wurde. Nach Touren durch ganz Deutschland (inklusive Auftritt beim Kaiser) konnte ein Psychologe namens Oskar Pfungst schließlich nachweisen, dass Hans die erwünschten Antworten an unbewussten Körpersignalen der Fragesteller ablas (Mlodinow, 2013). „Hans mag kein Pferde-Einstein gewesen sein, aber er verfügte über ein empfindliches Wahrnehmungsvermögen, das ihm erlaubte, subtilste Reaktionen zu interpretieren. Zudem bilden Pferde abstrakte Konzepte, kennen ihren Namen und sind mitunter sogar in der Lage, Werkzeuge benutzen.“ (Breuer, 2012).

Analog zu diesem Fall sollten wir uns nicht von oberflächlichen Qualitäten des Programm-Outputs leiten lassen, wenn wir die Fähigkeiten einer KI wie ChatGPT beurteilen wollen. Wer kritisch über das Verhältnis von Geist und Sprache nachdenkt, kommt zudem nicht an Ludwig Wittgenstein vorbei. Ihm zufolge liegt die Bedeutung von Wörtern weder in irgendwelchen „Gegenständen“ noch in unseren „Gedanken“. So sagen wir z.B., dass „Zwei“ der Name einer Zahl oder „Westen“ der Name einer Himmelsrichtung ist. Aber, wenn wir von der Rolle dieser Wörter im Satz als Subjekte auf ihre Wirklichkeit schließen, tapen wir in eine Falle. Denn weder „Zwei“ noch „Westen“ existieren in diesem Sinne (Wittgenstein, 1984, §19, 23 u. 133). In Bezug auf mentale Attribute wie „Denken“ oder „Sprechen“ tapen wir in eine ähnliche Falle, wenn wir sagen, dass eine KI wie ChatGPT „denkt“.

Denken ist kein Vorgang oder eine formale Funktion, die von einer Maschine ausgeführt werden kann. Vielmehr zeigt das Nachdenken über Denken, dass vielfältige Verhaltensweisen zu ihm gehören. Beispiele hierfür sind „kritisch sein“, „schlussfolgern“, „etwas glauben“, „gedankenlos handeln“, „fantasieren“, „etwas einschätzen“ oder „einen Geistesblitz haben“. Eine ähnliche Liste lässt sich für „sprechen“ bzw. „Sprache“ aufstellen, die vom Selbstgespräch, Witze machen und Kommandieren über das Diskutieren bis hin zum Theaterspiel reicht. Der Punkt ist, dass es sich bei Fähigkeiten wie Denken oder Sprechen um „weit verzweigte Begriffe“ handelt, deren Gebrauch in diejenigen Situationen eingebettet ist, die unser Leben ausmachen (Hacker, 2013). In der menschlichen Sprache spiegelt sich somit nichts weniger als unsere „Lebensform“ wider. Ein Programm wie ChatGPT teilt diese Lebensform aber nicht mit uns. Denn ein Computer hat keine Ziele,

Präferenzen, Motivation, Gefühle oder Erfahrungen. Er verfügt auch nicht über das körperliche Ausdrucksvermögen, das einen wichtigen Teil des Denkens bildet. Denken Sie hier z.B. einmal an die Skulptur „Der Denker“ von Rodin! Wie zum Beispiel der Philosophieprofessor Gunnar Schumann herausstellt, ist Sprache „keine individuelle Weltansicht, sondern nur möglich, wenn alle Sprecher*innen im Prinzip dieselbe Weltansicht haben“ (2018, S. 498). ChatGPT und Computer im Allgemeinen teilen unsere Weltansicht jedoch nicht (Darstellung nach Fesefeldt, 2022; aber auch Otte, 2023, stellt eine Überschätzung der Fähigkeiten von KI fest).

Das Modelltraining eingerechnet, verbraucht eine einzige Anfrage (Prompt) an ChatGPT ungefähr so viel Strom wie eine 5-Watt-Glühlampe in über einer Stunde und hat zudem weitere Kosten. Im Gegensatz dazu kommt das gesamte menschliche Gehirn mit dem Energieverbrauch einer 20-Watt-Glühlampe aus. Trotzdem erreicht es damit wohl ein Viertel der nominellen Rechengeschwindigkeit der heutzutage weltweit schnellsten Supercomputer. Einer von ihnen ist der Fugaku in Kobe (Japan). Er erzielt eine Rechengeschwindigkeit von ca. 450.000 Teraflops bei einem Energieverbrauch von ca. 30 Millionen Watt. Das Gehirn mit seinen ca. 86 Milliarden Neuronen erreicht schätzungsweise 10.000 bis 100.000 Teraflops bei einem Energieverbrauch von 20-30 Watt.

Da unser Gehirn in puncto Effizienz und Vielseitigkeit immer noch unschlagbar ist, sollten wir uns gut mit seiner „Bedienungsanleitung“ vertraut und von ihm Gebrauch machen (Hüther, 2016). Ihr Gehirn wird besonders gefragt sein, wenn es darum geht, die Stärken und Schwächen von kommenden KI-Systemen kritisch zu hinterfragen, sie

sinnvoll anzuwenden und ihre Ergebnisse richtig einzuschätzen.

Literatur

- Breuer, H. (2012). *Tierische Intelligenz: Pferde sind viel klüger als gedacht*. Online unter: <https://www.tagesspiegel.de/wissen/pferde-sind-viel-klueger-als-gedacht-6412090.html> (letzter Zugriff am 04.12.2023)
- ByteByteGo (2023). *How ChatGPT Works Technically I ChatGPT Architecture*. Online auf Youtube unter: <https://www.youtube.com/watch?v=bSVTVREwSNw> (letzter Zugriff am 28.11.2023)
- Ertel, W. (2016). *Grundkurs Künstliche Intelligenz. Eine praxisorientierte Einführung*. Wiesbaden: Springer.
- Epstein, R., Roberts, G. & Bebbler, G. (2009). *Parsing the Turing Test. Philosophical Issues in the Quest for the Thinking Computer*. Wiesbaden: Springer.
- Fesefeldt, J. (2022). *Können Maschinen Denken? Künstliche Intelligenz im Spiegel von Wittgensteins Philosophie des Geistes* (Masterarbeit an der Fernuniversität in Hagen).
- Hacker, P.M.S. (2013). *The Intellectual Powers. A Study of Human Nature* (Kapitel 10). Chichester: Wiley-Blackwell.
- Hüther, G. (2016). *Bedienungsanleitung für ein menschliches Gehirn*. Göttingen: Vandenhoeck & Ruprecht.
- Kissinger, H., Schmidt, E. & Huttenlocher, D. (2022). *The Age of AI*. London: John Murray Publishers.
- Krohn, J., Beyleveld, G. & Bassens, A. (2019). *Deep Learning illustriert*. Heidelberg: dpunkt.verlag.
- Milmo, D. (2023). ChatGPT reaches 100 million users two months after launch. *The Guardian* (Artikel in der Onlineausgabe vom 02.02.2023).
- Modinow, L. (2013). *Subliminal: How Your Unconscious Mind Rules Your Behavior: How Your Unconscious Mind Rules Your Behavior*. Knopf Doubleday Publishing Group.
- OpenAI – Research Index (2023). Website online unter: <https://openai.com/research> (letzter Zugriff am 24.11.2023)
- Otte, R. (2023). Die künstliche Intelligenz wird gerade gewaltig überschätzt. *Frankfurter Allgemeine Zeitung*, Onlineausgabe vom 18.06.2023.
- Patzig, G. (Hrsg.). (2008). *Gottlob Frege - Funktion, Begriff, Bedeutung. Fünf logische Studien*. Göttingen: Vandenhoeck & Ruprecht.
- Schumann, G. (2018). Sprachlicher Kulturrelativismus oder Universalismus. In Busche, H., Heinze, T., Hildebrandt, F. & Schäfer, F. (Hrsg.), *Kultur – Interdisziplinäre Zugänge*. Wiesbaden: Springer.
- Stephan, A. & Walter, S. (Hrsg.). (2021). *Computing Machinery and Intelligence. Können Maschinen denken?* (enthält Turings Aufsatz von 1950). Ditzingen: Reclam.
- Weitz, E. (2023). *ChatGPT und die Logik*. Online auf Youtube unter: https://www.youtube.com/watch?v=5cYyewwYF_0 (letzter Zugriff am 23.11.2023)

Wittgenstein, L. (1984). *Werkausgabe in 8 Bänden - Band 1: Philosophische Untersuchungen*. Berlin: Suhrkamp (Erstauflage der Phil. Untersuchungen 1953).

Wohlers, C. (Hrsg.). (2009). *René Descartes – Meditationen*. Hamburg: Felix Meiner (enthält Descartes "meditationes de prima philosophia" von 1641).

Kontakt

Johannes Fesefeldt

Diplom-Psychologe &
M.A. Philosophie
Deutsche Gesellschaft für
Personalwesen e. V.



Johannes Fesefeldt ist am Standort Berlin als Psychologe tätig. Seine Schwerpunkte liegen in der Personalauswahl nach den Standards der DIN 33430 und der Personalentwicklung. Dazu gehört die Konzeption und Moderation von Assessment Centers für Führungspositionen im öffentlichen Dienst, psychologische Beratung bei den Auswahlverfahren des Auswärtigen Amtes sowie auch Trainingsseminare zu den Themen Teammanagement und Kommunikation für Führungskräfte. Darüber hinaus beschäftigt er sich mit den Potenzialen von künstlicher Intelligenz.

✉ fesefeldt@dgp.de