Was HR-Verantwortliche über KI-Sprachverarbeitung wissen sollten

Der Einsatz von künstlicher Intelligenz (KI) in Organisationen verspricht Wunderdinge. Die Realität sieht noch anders aus, auch in HR-Abteilungen. Doch was können Personalverantwortliche von Sprachmodellen wie Chatbots erwarten? Zunächst einmal, dass sie es bei KI nicht mit einem technischen Ebenbild des Menschen zu tun haben. Die Metaphorik, die den „Kollegen Computer" umgibt, verhüllt seinen wahren Charakter.

Seit etwa 2018 rückt der Einsatz künstlicher Intelligenz zunehmend ins Blickfeld der Personalarbeit. Zunächst sorgten zahlreiche Studien und Fachbeiträge für große Erwartungen – oft begleitet von einer gewissen Euphorie. Die Begeisterung ist nicht unbegründet, doch die tatsächliche Integration von KI in HR-Prozesse verläuft weniger rasant als zunächst angenommen. Zum Beispiel nutzen erst circa zehn Prozent der deutschen Personalabteilungen KI im Regelbetrieb. Insgesamt scheinen vor allem wirtschaftliche, rechtliche und ethische Bedenken sowie ein erlebtes Kompetenzdefizit die KI-Transformation zu hemmen (Fesefeldt, 2025).

Das Hauptanwendungsfeld von KI, nicht nur im Personalwesen, liegt in der natürlichen Sprachverarbeitung auf Basis sogenannter Large Language Models (LLMs). Zu denen gehört zum Beispiel das Modell GPT, entwickelt vom US-Software-Unternehmen OpenAI, oder Gemini von Google. Anwendungen, die auf den LLMs aufsetzen, sind zum Beispiel sogenannte Chatbots, etwa ChatGPT oder Microsoft Copilot; beide basieren auf GPT. Von der Hilfe beim Onboarding bis hin zum Führungsfeedback durch KI scheint es unbegrenzte Einsatzmöglichkeiten zu geben. Doch ein Vergleich mit unserer eigenen Intelligenz zeigt fundamentale Unterschiede von LLMs und heutigen künstlichen Intelligenzen im Allgemeinen auf. Wir werfen in diesem Beitrag zunächst einen Blick auf diese, um dann Vorteile und Grenzen der KI zu beleuchten. Zuletzt gehen wir darauf ein, wie man gelungen mit KI „kommuniziert" und wie die Rolle von Menschen in der „Zusammenarbeit" mit KI aussehen sollte.

Verstehen wir uns?

Der griechische Philosoph Aristoteles (384 bis 322 v. Chr.) bezeichnete den Menschen als „das sprechende Wesen", dessen Kommunikationsfähigkeit das Zeichen von „Vernunft" sei. Sprache gilt also seit jeher als zentrales Merkmal menschlicher Intelligenz. Jedoch ermöglichte erst die Erfindung des Computers die Konstruktion von „sprechender" KI. Während die ersten Programme (Chatbots) in den 1960er-Jahren geschrieben wurden, reflektierte der britische Mathematiker Alan Turing bereits 1950 ihre Folgen. In seinem Aufsatz „Computing Machinery and Intelligence" schlug das Computergenie den weltberühmten Turing-Test vor.

Ziel des Tests war es, zu überprüfen, ob eine Maschine während einer fünfminütigen Gesprächssequenz in der Lage ist, menschliche Kommunikation so gut zu imitieren, dass ein Proband sie nicht mehr von einem echten Menschen unterscheiden

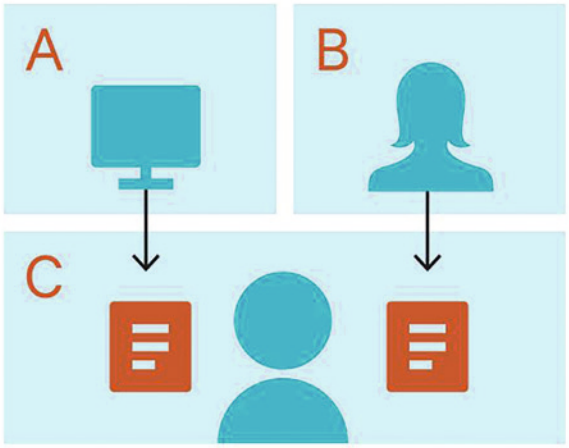

Abbildung 1: Originaler Aufbau des Turing-Tests; A ist ein Computer beziehungsweise eine KI (eigene Darstellung).

Foto: StockPhotoPro/AdobeStock

„Erst circa zehn Prozent der deutschen Personalabteilungen nutzen KI im Regelbetrieb.“

kann. Der Test wurde darum auch als „Imitation Game“ bezeichnet (siehe Abbildung 1).

Als intelligent sollte eine Maschine gelten, wenn es ihr in mindestens 30 Prozent der Durchgänge gelänge, einem Menschen vorzutäuschen, selbst ein Mensch zu sein. Turing prognostizierte, dass um das Jahr 2000 herum künstliche Intelligenzen existieren würden, die den Turing-Test bestehen. Tatsächlich explodierte dann um die 2000er-Jahre herum das sogenannte *Deep Learning* in künstlichen neuronalen Netzwerken (KNN), deren Lernalgorithmen der Funktion des menschlichen Gehirns nachempfunden sind. In vielen Bereichen wurden KNN zum vorherrschenden KI-Ansatz. Big Data und Cloud-Computing in weltweit verteilten, teils über 100 Hektar großen Datenzentren dienten als „Enabler“. Erstmals kam KI in die Nähe von menschlichen Intelligenzleistungen oder überbot sie sogar, so etwa bei der Bilderkennungsfehlerrate. Allerdings erschien das Bestehen des Turing-Tests immer noch als gordischer Knoten, den kein KI-Chatbot durchschlagen konnte.

Nun gaben Forscher der University of California, San Diego, Anfang 2025 bekannt, dass OpenAIs ChatGPT (Version 4.5) einen echten Turing-Test bestanden habe. Die Studie von Jones und Bergen wurde breit rezipiert und diskutiert, hat allerdings noch keine Peer-Review durchlaufen. Turing hatte mit seinem Test die tieferen, philosophischen Fragen nach der Natur des Denkens, der Sprache oder des Bewusstseins beiseiteschieben wollen, um die KI-Entwicklung zu fördern, aber das Bestehen eines Turing-Tests wirft die Frage auf, ob der Output beziehungsweise das „Verhalten“ einer KI wirklich der einzige Maßstab für ihre Intelligenz sein sollte. Turing selbst war der Überzeugung gewesen, dass wir uns als Menschen gegenseitig Fähigkeiten wie „Denken“, „Sprache“ oder „Bewusstsein“ bloß aufgrund des damit einhergehenden äußeren Verhaltens zuschrieben. Wenn nun auch eine Maschine dieses Verhalten an den Tag legt, warum sollten wir diese „höfliche Übereinkunft“ irgendwann nicht für sie gelten lassen?

Wie Large Language Models funktionieren

Schauen wir uns zuerst die Funktionsweise von LLMs etwas näher an: Als Trainingsdaten erhalten die Modelle „Weltwissen“ aus frei zugänglichen Internetseiten, Artikeln, Büchern, Foren, Eingaben der Nutzer*innen und vielem mehr; inzwischen können zumindest die großen Agenten auch mit Audio-, Bild- oder Videodaten umgehen. Der KI-Algorithmus verknüpft diese Daten mithilfe von Wahrscheinlichkeitsrechnungen. Die Daten werden dazu zunächst in sogenannte *Tokens* zerlegt, die (aus menschlicher Sicht mal mehr, mal weniger verständliche) Bestandteile von Wörtern und Wortfolgen repräsentieren. Beispielsweise lässt sich „Kraftfahrzeug“ (intuitiv) in „Kraft“, „fahr“ und „zeug“ zerlegen. Für jeden Token kann die Position im semantischen Raum des LLM exakt angegeben werden, weil die Wahrscheinlichkeitsbeziehungen zu sämtlichen anderen Tokens des LLM im Vektor des Tokens gespeichert werden (siehe Abbildung 2).

Gibt eine Nutzerin oder ein Nutzer bei ChatGPT zum Beispiel den Prompt „Zwei plus zwei …“ ein, sucht der Algorithmus den wahrscheinlichsten nächsten Token, in diesem Fall zuerst den Token „ist“, anschließend den Token „gleich“ und schließlich den Token „vier“. Obwohl es im

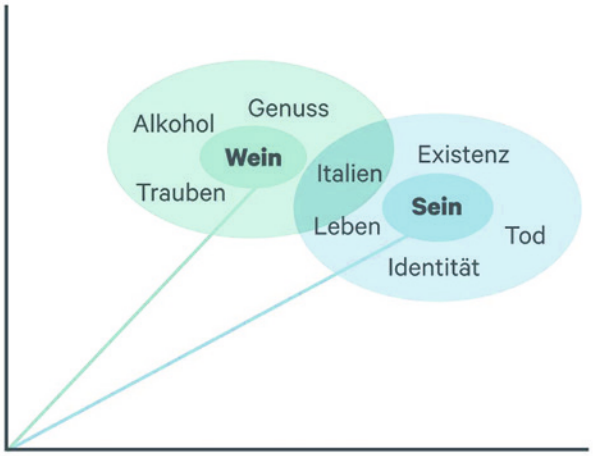

Abbildung 2: Repräsentation zweier Wörter als Vektoren im semantischen Raum des LLM (eigene Darstellung).

Bereich des Deep Learning verschiedene Trainingsmethoden gibt, ist das Training der KI stets von menschlichen Zielvorgaben in den Daten abhängig. Beispielsweise wurden beim Training von GPT die von GPT generierten Antworten zunächst sämtlich und in einer weiteren Trainingsphase zumindest stichprobenartig von Menschen bewertet.

Die Grenzen von Large Language Models

Aus der Funktionsweise von LLMs ergeben sich ihre Schwächen und Risiken: Der Algorithmus hat kein Konzept von „richtig" oder „falsch", sondern folgt allein einem Wahrscheinlichkeitsmechanismus zur Auswahl einer Antwort. Wenn eine Aussage oft in vertrauenswürdigen Quellen vorkam, ist zwar die Wahrscheinlichkeit hoch, dass das LLM sie korrekt wiedergibt. Aber: Das Modell kann auch falsche oder veraltete Informationen wiederholen, wenn diese in den Trainingsdaten enthalten waren. Landläufig spricht man dann von „Halluzinationen" der KI. Einige LLMs sind daher mit externen Wissensdatenbanken oder Suchsystemen verbunden (sogenannte *Retrieval-Augmented Generation*). Dabei wird eine Suchanfrage an einen speziellen Datensatz oder an Webquellen gestellt, um thematisch belastbare Informationen zu erhalten. Auch gibt es speziell auf die Faktenprüfung trainierte LLMs.

Im August 2024 wurde mit dem „EU AI Act" das weltweit erste umfassende Gesetz zum Umgang mit KI verabschiedet. Zwar lässt sich auch die europäische Datenschutzgrundverordnung von 2016 auf KI-Systeme beziehen, sie legt den Fokus aber auf die Verarbeitung von personenbezogenen Daten. Der AI Act, der erst Mitte 2027 vollumfänglich in Kraft tritt, führt unter anderem ein Risikostufensystem für KI ein. Zum Beispiel stellt ein KI-gestütztes Bewerbungsverfahren ein „hohes Risiko" dar. Doch bereits der Einsatz eines KI-Chatbots wie ChatGPT im Kundenservice („begrenztes Risiko") muss den Nutzer*innen offengelegt werden. Im Bereich der Hochrisikosysteme, die sich durch eine Gefahr für die Grundrechte von Personen auszeichnen, gelten hohe Anforderungen. Insbesondere ist vor dem ersten Einsatz der KI eine „Grundrechtsfolgenabschätzung" nötig, die auch Maßnahmen zur Risikominimierung enthält.

Die EU hat damit umfangreich auf die Schwächen von KI reagiert, zu denen auch die Probleme *Black Box* und *Bias* gehören. Mit Bias wird allgemein die Tendenz von KI-Algorithmen bezeichnet, falsche oder einseitige Informationen zu produzieren, insbesondere wenn dadurch bestimmte Personengruppen benachteiligt werden. Die Folge ist eine voreingenommene KI, die gesellschaftliche Vorurteile, Stereotype und Diskriminierungen erlernt hat und möglicherweise noch verstärkt. Weil das Lernen in den LLMs grundsätzlich nicht nachvollziehbar ist, bleibt der Bias womöglich verborgen. Ein Beispiel aus dem Bereich der generativen KI: Bittet man eine KI, die mit Bildern weißer CEOs gefüttert wurde: „Show me an image of a CEO", erhält man natürlich das Bild eines weißen, männlichen CEO. Im Fall des chinesischen LLM DeepSeek, das wegen seiner hohen Kosteneffizienz Aufsehen erregte, tritt zudem das Problem der Zensur bestimmter (politischer, historischer, aber auch anderer) Inhalte auf den Plan.

Es gibt zum Glück mehrere Strategien, um dem Bias zu begegnen: Wenn die KI mit diversen Daten trainiert wird, sinkt der Bias. Das können etwa Daten sein, die verschiedene Personengruppen umfassen. Er sinkt auch, wenn der Algorithmus durch menschliche Kontrolle des Outputs und durch Feedback korrigiert wird, und zwar bestenfalls bereits in der Trainingsphase. Auch hier gilt aber, dass die KI letztlich nur auf Wahrscheinlichkeiten reagiert und kein wirkliches Verständnis ihres Bias erlangt. Auf Seite der Nutzer*innen liegt ein Schlüssel in der passenden Auswahl der KI. Zum Beispiel ist das LLM Claude des amerikanischen Unternehmens Anthropic bei logikbasierten Aufgaben stärker als GPT, aber schwächer bei sprachlichen. Ein weiterer Schlüssel ist das sogenannte *Prompting*, das heißt die gelungene „Kommunikation" mit der KI.

„Der Algorithmus hat kein Konzept von ‚richtig' oder ‚falsch', sondern folgt allein einem Wahrscheinlichkeitsmechanismus zur Auswahl einer Antwort."

Vom guten Prompting

Advanced Prompting zeichnet sich gegenüber dem einfachen „Darauflosfragen" durch die Beachtung verschiedener Qualitätskriterien aus: Der Prompt sollte klar und möglichst wenig mehrdeutig formuliert sein, genügend Hintergrundinformationen und eine Vorgabe für die Struktur der Antwort (etwa Fließtext, Liste oder Tabelle) sowie eine Zielgruppe (zum Beispiel Expert*innen oder Laien) enthalten. Grundsätzlich gilt: Je mehr solcher „Variablen" in das Prompting einfließen, desto präziser wird die Antwort der KI ausfallen. Um das beste Ergebnis zu finden, müssen in der Regel mehrere Schleifen (Iterationen) durchlaufen werden, in denen Variablen verändert oder ergänzt und die Outputs miteinander verglichen werden.

Um Halluzinationen zu vermeiden und generell bessere Ergebnisse zu bekommen, kann die KI aufgefordert werden, eine bestimmte Rolle einzunehmen, zum Beispiel: „Du bist nun eine Expertin für Personalpsychologie." Mithilfe von *Chain-of-Thought Prompting* können komplexe „Gedankengänge" der KI schrittweise sichtbar gemacht und strukturiert werden. Auch sollten dieselben oder ähnliche Prompts wiederholt daraufhin geprüft werden, ob sie dieselben inhaltlichen Ergebnisse produzieren (Konsistenz-Check). Außerdem ist es oft sinnvoll, ein neues Chatfenster zu öffnen und dort das Prompting gegebenenfalls auf Basis einer Zusammenfassung der bisherigen Ergebnisse neu zu starten, um das einige Tausend Zeichen umfassende „Gedächtnis" des LLM zu leeren. Die Aktivierung der erwähnten Retrieval-Augmented Generation erlaubt zudem den Zugriff auf themenspezifisch trainierte Modelle. Generell sollte die KI zusätzlich gebeten werden, ihre Quellen anzugeben. Für den Fall, dass es keine gibt oder diese unklar sind, muss die KI explizit aufgefordert werden, dies zuzugeben.

Ethisches Prompting geht noch einen Schritt weiter. Idealerweise wird das Modell vorab für seinen Einsatzzweck „getuned", indem es mit neuen, repräsentativen Daten gefüttert wird, zum Beispiel die Daten von Männern und Frauen betreffend. Auch

sollten alle Prompts sensibel formuliert werden, um die Aktivierung oder Verstärkung von Bias zu vermeiden. Die KI kann zusätzlich darum gebeten werden, das Thema aus möglichst unterschiedlichen Blickwinkeln zu analysieren, etwa aus verschiedenen kulturellen oder geschlechtsbezogenen Perspektiven, was einseitige Ergebnisse verhindert. Außerdem kann sie noch einen Auftrag erhalten wie etwa: „Reflektiere deine Antwort selbstkritisch." Bei alldem bleiben die menschliche Reflexion und Kontrolle des Outputs trotzdem unerlässlich.

Fazit

Denken ist kein formaler Vorgang, der einfach von einer Maschine reproduziert werden könnte. Vielmehr ist es, wie man sprachphilosophisch sagen kann, ein „weitverzweigter Begriff", der Verhaltensweisen wie „kritisch denken", „nachdenken", „schlussfolgern", „sich erinnern" oder „einen Geistesblitz haben" beinhaltet. Noch grundsätzlicher kann man argumentieren, dass Computer nicht an unserer Lebensform teilnehmen. Ihre Formen von „Intelligenz" und „Sprache" entspringen nicht dem Leben. Ein Computer hat keine wirklichen Gedanken, Erinnerungen, Gefühle oder Motive, nicht einmal Präferenzen.

KI wird unsere Intelligenz, Erfahrung, kritische Reflexion und unser Fachwissen zumindest mittelfristig nicht ersetzen können, aber als Suchmaschine, Inspirationsquelle und digitales Helferlein eine beeindruckende „Karriere" in der Personalarbeit machen, und zwar in allen Funktionen. Die scharfe Grenze zwischen natürlicher und künstlicher Intelligenz erkennen wir zum Beispiel auch daran, dass bereits ein einziger Prompt an GPT so viel Energie verbraucht wie unser gesamtes Gehirn in einer Viertelstunde (circa fünf Watt). Mit unserem Hirn erreichen wir aber eine geschätzte Rechengeschwindigkeit von 10^{15} bis 10^{25} Rechenoperationen pro Sekunde, was etwa einem Viertel der schnellsten Supercomputer entspricht. Die Kraft unseres Gehirns ist also gewaltig. Wir sollten sie nutzen, um in ihr die Bedienungsanleitung für KI zu finden!

LITERATUR:

European Union. (2025). Artificial Intelligence Act: Regulation (EU) 2024/0138 laying down harmonised rules on artificial intelligence and amending certain Union legislative acts. *Official Journal of the European Union*, Z C/2025/1027. Verfügbar unter eur-lex.europa.eu/eli/C/2025/1027/oj/eng

Fesefeldt, J. (2024). *Warum ChatGPT und andere KIs so „dumm" sind wie ein Pferd.* Deutsche Gesellschaft für Personalwesen e. V. Verfügbar unter www.dgp.de/chatgpt-intelligenz

Fesefeldt, J. (2025). Der Status Quo von KI im Personalwesen. *dgp-Informationen.* Verfügbar unter www.dgp.de/veroeffentlichungen/dgp_informationen

Fessler, R., Toklu, A., Behnke, Y., Pfiel, U. & Bolecek, R. (2023). *Künstliche Intelligenz für Unternehmer.* Zürich: Mensch Verlag.

Jones, C. R. & Bergen, B. K. (2025). Large Language Models Pass the Turing Test. *arXiv.org.* doi.org/10.48550/arXiv.2503.23674

Turing, A. M. (1950). Computing Machinery and Intelligence. *Mind, 59*(236), S. 433–460. doi.org/10.1093/mind/LIX.236.433.

DIE AUTOR*INNEN:

Johannes Fesefeldt
Diplom-Psychologe, M. A. Philosophie. Berater der Deutschen Gesellschaft für Personalwesen (dgp) am Standort Berlin. KI im Personalwesen ist einer seiner Themenschwerpunkte.

fesefeldt@dgp.de

Dr. Anna-Lena Jobmann
Diplom-Psychologin. Sie entwickelt seit 2016 für die dgp Eignungstests und berät den öffentlichen Dienst unter anderem zu datenbasierten Personalentscheidungen.

jobmann@dgp.de

Fotos: privat (oben), dgp (unten)

Spannende Themen, **hochrelevant**!

WIRTSCHAFTS-PSYCHOLOGIE aktuell

Jetzt den Newsletter abonnieren, kostenlos!

wirtschaftspsychologie-aktuell.de/newsletter

www.wirtschaftspsychologie-aktuell.de

3 | 2025

WIRTSCHAFTS-PSYCHOLOGIE aktuell

Zeitschrift für Personal und Management

ONBOARDING

Glück bei der Arbeit

Aktuelle Studien zeigen: Erleben wir Arbeitsglück, gelingt uns der Start in einem neuen Job leichter

Weltanschauung

Wie Menschen die Welt sehen, kann auch ihr Verhalten im Job prägen

Falsche Anreize

Was Unternehmen bei der Mitarbeitendenmotivation besser machen können

Was kann KI?

Was ist von digitalen HR-Assistenten zu erwarten? Eine Einführung