Gerhard Lauer

Das Selbst der Künstlichen Intelligenz

Der Aufsatz geht der Frage nach, wie wahrscheinlich ist es, dass Künstliche Intelligenz mehr als ein Assistenzsystem im Kontext der psychotherapeutischen Arbeit werden könnte. Zur Abschätzung der Plausibilität einer Antwort diskutiert der Beitrag die Konzeption der derzeit verfügbaren großen KI-Sprachmodelle und versucht kritisch die jüngsten Entwicklungen dieser Modelle in die Zukunft zu projizieren. Im Ergebnis kommt der Aufsatz zu dem Schluss, dass nach derzeitigem Kenntnisstand ein künstliches Psychotherapeutensystem noch in einiger Ferne liegt, aber die Psychotherapie gut daran tun würde, sich in die Diskussion um eine komplementäre Intelligenz einzubringen.

1. Betrachtungen im Spiegel des Selbst

Künstliche Intelligenz erkennt sich nicht selbst im Spiegel. Das können außer uns Menschen ansonsten nicht nur viele andere Primaten, sondern auch Elefanten, Elstern und sogar Putzlippfische, um nur einige Arten zu nennen, bei denen diese Fähigkeit nachgewiesen ist. Zwar antworten inzwischen die multimodale Künstliche Intelligenz-System stimmig auf die Frage, warum sie sich nicht in einem Spiegel erkennen können, – sie seien nur auf Sprache und Texten, Bilder und Videos trainiert und daher ohne Erfahrung, Bewusstsein oder Emotion –, aber ein Selbst haben sie nicht. Eine Wahrnehmung als ein erkennendes Subjekt fehlt ihnen. Anders gesagt, können Künstliche Intelligenzsysteme nicht Objekt ihrer eigenen Erfahrung sein. Sie sind in der klassischen Formulierung von William James in seinen „Principles of Psychology" von 1890 kein ‚Me', aber auch kein ‚I'. Mit dieser Unterscheidung in ‚Me' und ‚I' hat 1890 William James eine der Grundfragen der von ihm mitbegründeten Psychologie versucht, einer Antwort näherzubringen, der Frage nämlich, wer ich bin und woher ich das weiß, wer ich bin. Nach James beschreibt das ‚I'-Selbst das akute Selbsterleben, das Selbst als Subjekt, während das ‚Me'-Selbst die Objektivierung meiner selbst bezeichnet, jene besonders bei Menschen ausgeprägte Fähigkeit, mich selbst zu betrachten, ob im Rückblick, in der gegenwärtigen Situation oder auch im Ausblick. Erst das Zusammenspiel von ‚I' und ‚Me' ermöglicht uns, dass wir uns als ein

 https://doi.org/10.1515/9783111509341-013

mehr oder minder konstantes Ich wahrnehmen. Die Entwicklung einer selbstbezogenen Wahrnehmung, Einschätzung und eines Wissens ist die Bedingung für den Aufbau eines Selbst. Mit den nicht wenigen Störungen bei dieser Entwicklung und Aufrechterhaltung des Selbst über die Lebensspanne hinweg hat es die Psychotherapie wesentlich zu tun.

Künstliche Intelligenzsysteme verfügen über keine solche Struktur des Verhältnisses von ‚I'-Selbst und ‚Me'-Selbst und damit über kein ganzheitliches Ich. Sie können daher auch nicht Patient*in einer Psychotherapie sein. In der Theorie William James' fehlt KI die Selbstwahrnehmung als ein materielles Selbst, also die Erfahrungen des eigenen Körpers. Zweitens mangelt der KI die Wahrnehmung als ein soziales Selbst, das aus dem Bewusstsein des eigenen Ansehens bei anderen erwächst und daher auch die Normen und kulturellen Werte des sozialen Miteinanders umfasst. Drittens verfügt KI auch nicht über ein spirituelles Selbst, welches die Erfahrungen der eigenen seelischen Vorgänge beinhaltet wie Einstellungen, Glauben und Gedanken einer Person. Obgleich sich KI in Verhaltenstests wie dem Turing-Test verhält, wie es ein Mensch tun würde (Mei et al., 2024) ist es nach James' Theorie KI subjektlos.

Neuere Theorien wie etwa die Theorie der freien Energie Karl Fristons and Christopher Friths (2015), die versuchen, die Funktionsweise des menschlichen Gehirns als probablistisches Vorwegnehmen möglicher Erfahrungen zu verstehen, kommen zu ähnlichen Schlüssen wie schon William James vor mehr als hundert Jahren. Aus dem Rauschen der neuronalen Signale muss das Gehirn aus letztlich physikalischen Gründen der Energiehaushaltung die für den Organismus wesentlichen Signale herausfiltern und gegebenenfalls in die bewusste Wahrnehmung heben, gerade dann, wenn die Voraussagen des Gehirns auf eine unerwartete Situation nicht mehr passen und nach neuen Lösungen gesucht werden muss. Um zugleich routinierte Voraussagen wie auch die Adaption an unvorhersehbare Zustände zu ermöglichen, funktioniert das menschliche Gehirn nicht als Reiz-Reaktion-Mechanismus, sondern testet zumeist in kleinem Umfang ständig auch andere wahrscheinliche Lösungen. Über einen bestimmten Grad des Rauschens oder dem zufälligen Springen von einem neuronalen Erregungsmuster zu einem anderen konkurrieren neuronale Mustern miteinander und darüber gewinnen Muster mehr oder minder wahrscheinlich eine Wirkung für die Selbstwahrnehmung. Die Selbstwahrnehmung hängt daher eng mit einem Hintergrundrauschen verschiedener neuronaler Muster zusammen und entsteht gerade nicht in einer strengen Abfolge entlang von Entscheidungsbäumen. Menschliche Intelligenz ist aufgrund

dieses komplexen-bayesschen Selbstverhältnisses eine andere Art von Intelligenz, als es Künstliche Intelligenz derzeit ist. Sie ist eine Intelligenz des Selbstverhältnisses, ja das Wort Selbst hat nur Sinn im Zusammenhang mit menschlicher Intelligenz. Das Selbst aber ist die Quelle für das Verstehen anderer, für die gezielte Aufmerksamkeit für andere, auch für deren Schwierigkeiten, für Sorge, Mitleid und Empathie. Jede Form der Psychotherapie hängt an dieser Besonderheit der nicht-künstlichen Intelligenz, ein Selbstverhältnis ausbilden zu können. Mit dieser ersten Auskunft mag man sich beruhigen und folgern, dass die Psychotherapie mit KI eine Helferin hat, mehr aber auch nicht. Diese KI-Helferin ist in der Psychotherapie bereits verschiedentlich etabliert (Ebert & Baumeister, 2023), wenn auch Details der Ausgestaltung von Rollen und Funktionen der KI für die psychotherapeutische Arbeit umstritten sind (Deutsche Gesellschaft für Psychologie, 2023; Abrams, 2024).

2. Die Temperatur der LLM

Doch die Unterscheidung in menschliche und künstliche Intelligenz ist nicht so klar, wie es zunächst scheint, denn Künstliche Intelligenz führt das Wort ‚Intelligenz' nicht nur metaphorisch im Namen (Wolfram Alpha, 2023). Einem intelligenten Organismus ähnlich lernen KI-Systeme Wörter in ihren wahrscheinlichen Kontexten einzuordnen. Trainiert auf Milliarden von Wort- und Satzkombinationen haben die großen Systeme ein fast akkurates Wissen davon, wie wahrscheinlich bestimmte Wortkombinationen sind, wenn sie durch einen Prompt befragt werden. Sie errechnen den Bedeutungsraum, in dem sie Sätze in Tokens einzelner Wörter zerlegen und messen, wie wahrscheinlich Nähe und Ferne anderer Wörter zu dem jeweiligen Token sind. Dann werden den Tokens entsprechende Definitionen zugewiesen, indem sie rechnerisch in einem Bedeutungsraum mit Wörtern mit ähnlicher Bedeutung angeordnet werden. Ein Wort wie „Psychotherapie" steht dann in einem Bedeutungsraum, dem Embedding, mit Wörtern wie „Patient" oder „Psychotherapeut". Sprachmodelle nutzen daher keine Grammatik, sondern Vorkommenshäufigkeiten, die über Milliarden von Trainingsrunden abgleicht, wie weit das Modell mit seinen Voraussagen tatsächlich existierenden Texten gleicht. Modelle verbessern sich so sukzessiv.

Für jedes Token im Vokabular des Modells hat das Aufmerksamkeitsnetz des Sprachmodells eine Wahrscheinlichkeit ermittelt, dass dieses Token das wahrscheinlichste ist, um als Nächstes in dem von ihm gene-

rierten Satz verwendet zu werden. Das Sprachmodell generiert ein Wort und gibt dann das Ergebnis an sich selbst zurück. Das erste Wort wird allein auf der Grundlage der Aufforderung erzeugt. Das zweite Wort wird dagegen schon generiert, indem das erste Wort in die Antwort einbezogen wird, dann das dritte Wort, indem die ersten beiden generierten Wörter einbezogen werden, und so weiter. Dieser Prozess, die sogenannte Autoregression, wird so lange wiederholt, bis die Aufgabe abgeschlossen ist. Oder in der Sprache der Computerwissenschaften formuliert, handelt es sich bei den gegenwärtigen KI-Sprachmodellen um Aufmerksamkeitsnetzwerke, die die kognitive Aufmerksamkeit des Menschen nachahmen, in dem sie für jedes Wort dessen Gewicht innerhalb eines Netzwerks von Wörtern vermessen. Diese Vermessung der Wortgewichte erfolgt dynamisch, ist also nicht eingefroren, und kann durch paralleles Prozessieren exponentiell wachsen. In der Summe bilden diese Netzwerke die Assoziationen zwischen Wörtern ab, wie sie Menschen tatsächlich nutzen. Damit bilden diese dynamisch errechneten Netzwerke der Wörter das menschliche Wissen ab, auf das sie mithilfe der Trainingsdaten kalibriert wurden. Die Systeme sind daher mindestens so intelligent, wie die Daten, mit denen sie trainiert wurden und wissen so auch, dass „Psychotherapie“ sachlich eng mit „Patient“ zusammengehört, aber auch mit „Krankenkasse“ oder auch mit „Gesundheit“ und anderen Worten, die dynamisch je nach Kontext unterschiedlich nahe zueinander stehen. So imitiert das künstliche Netzwerk das natürliche Netzwerk unseres Wissens. Von Intelligenz hier zu sprechen, ist keine bloße Metapher.

Die Sache mit der Voraussage des nächsten wahrscheinlichsten Wortes würde so allerdings zu immer ähnlichen Antworten führen und sich nicht menschlich anhören. Daher nutzen alle LLMs eine herabgestufte Wahrscheinlichkeit für das jeweils nächste Wort. Nicht das Wort, das am wahrscheinlichsten ist, wird gewählt, sondern ein etwas weniger wahrscheinliches Wort. Man spricht hier von der Temperatur eines Modells, das nicht auf 1.0 dem wahrscheinlichsten Wort ausgerichtet ist, sondern auf das etwas geringer wahrscheinliche mit der Temperatur von 0.8. Das macht die Antworten flexibler und kreativer. Für uns hören sich daher die Antworten menschlicher oder auch wärmer an. KI-Systeme sind also keine algorithmischen Maschinen, sondern ahmen Besonderheiten der menschlichen Intelligenz – deren probabilistisches Rauschen – gezielt nach.

Damit wird möglich, was niemand voraussehen konnte. Mit dem Anwachsen der Sprachmodelle erlangen die KI-Systeme emergente

Fähigkeiten (Wei, 2023; Ornes, 2023, dagegen Schaeffer, Miranda & Koyejo, 2023). Während die Sprachmodelle noch vor kurzem nur sehr kleine Dialoge führen konnten, dann schnell mit ihren Antworten im Kreis zu gehen begonnen haben und an Wissensdomänen gebunden blieben, bestehen Modelle wie Chat-GPT komplizierte Aufnahmetests etwa für ein Jura- oder ein Medizin-Studium und das besser als die meisten menschliche Kandidat*innen. Da die Modelle aufgrund ihrer Größe bzw. Tiefe nicht mehr von Menschen verstanden werden können, vermag niemand zu sagen, wo solche Schwellenwerte liegen, die ein größer werdendes Modell mit unvorhersagbaren Fähigkeiten ausstattet. KI-Systeme sprechen daher nicht nur immer besser die Sprache von uns Menschen, sondern gewinnen aus der Sprache auch Fähigkeiten, die sie vor kurzem noch nicht hatten. Sie erreichen nicht vorherzusagende Fähigkeiten etwa in der Bewältigung von komplexen Schlussfolgerungen und können sich auf nicht vorhersehbare Situationen immer besser einstellen (Pavlus, 2019; Meincke et al., 2024). Schon die Weiterentwicklung von ChatGPT 3.5 zu ChatGPT 4.0 innerhalb weniger Monate im Jahr 2023 zeigt, dass der Chatbot inzwischen Texte erstellen kann, die Fähigkeiten und Themen in einer Weise kombiniert, die nicht in den Trainingsdaten schon vorgekommen sind (Yu et al., 2023; Ananthaswamy, 2024). Die Charakterisierung der KI-System wie ChatGPT als stochastische Papageien (Bender et al., 2021) ist irreführend. Auch darin ähneln die KI-Systeme der menschlichen Intelligenz schon weit mehr als es in der öffentlichen Diskussion deutlich wird, auch wenn sie noch keine kontrafaktischen Analogieschlüsse ziehen können, wie das schon Kinder können (Lewis & Mitchell 2024).

Das Größenwachstum der Systeme hat allerdings zwei Seiten. Weil Speicher so viel günstiger geworden sind, können die Sprachmodelle überhaupt so groß werden, dass in einem Modell wie ChatGPT-3 nur 2048 Tokens zugleich prozessieren konnte, im GPT-4-Modell hingegen bereits Eingaben mit einer Länge von bis zu 32.000 Token verarbeitet werden können. Der Unterschied ist erheblich. Denn je mehr Text das Modell zugleich verarbeiten kann, desto mehr Kontext kann es erkennen, und desto besser werden seine Antworten auch auf unvorhergesehene Fragen sein. Das System hat durch den verbesserten Speicher mehr Wissen erworben, kann damit Wissen aus unterschiedlichen Domänen stimmiger rekombinieren und sich flexibler auf wechselnde Kontexte einstellen (Achiam et al., 2023). In fast allen Test wie etwa Aufnahmetests für Universitäten schlagen die KI-Systeme inzwischen Menschen. Doch das Wachstum hat einen Haken: Die erforderliche Re-

chenleistung steigt nicht linear mit der Länge der verarbeiteten Tokens, sondern exponentiell. Entsprechend steigt die Rechenleistung und zugleich steigen die Kosten dramatisch. Und noch etwas verkompliziert die Lage: Die Menge der Texte zum Training der Modelle wächst nicht mit (Villalobos et al., 2022). Auch wenn die neuesten Modelle auf unvorstellbar großen Textmengen trainiert wurden, wird das Trainingsmaterial inzwischen knapp (Knight, 2023). Rechtliche Auseinandersetzungen über die Legitimität der ohne Wissen ihrer Autor*innen für Trainingszwecke genutzten Texte und inzwischen auch Bilder für Large Multimodal Models (LMM) verknappen die Wachstumsmöglichkeiten noch einmal. Das Wachstum der Großen Sprachmodelle ist an eine natürliche Grenze gestoßen. Diskussionen um neue, etwa sogenannte BabyLMs oder auch Small Language Models (SLM) machen die Runde, mit denen Maschinen ähnlich wie Kinder Sprache an einer viel kleinen, dafür sehr viel besser kuratierten Datenmenge lernen sollen, um menschlich sich anfühlende Antworten zu generieren (BabyLM, 2023; Javaheripi & Bubeck, 2023). Die Zeit der großen Sprachmodelle ist vermutlich schon vorbei. Die Modelle nur einfach größer zu konzipieren, reicht nicht im Rennen um die Erstellung einer Künstlichen Intelligenz, andere Wege sind gefragt.

Während KI für immer mehr Lebensbereichen eine selbstverständliche Hintergrundtechnologie bildet, zugleich die Öffentlichkeit über KI noch staunt und nach einer Einordnung der jüngsten Entwicklungen sucht, sind die Entwickler schon dabei neue Wege zu suchen, die Artificial Intelligence (AI) tatsächlich in eine Artificial General Intelligence (AGI) zu heben. Dazu muss die KI verkörpert werden und Emotionen passend zu dem geführten Dialog ausdrücken können (Mishra, 2023). Embodied-AI-Systeme wie PALM-E von Google beispielsweise werden nicht nur mit Textdaten, sondern auch mit sensorischen Daten trainiert, um Roboter flexibel für unterschiedliche, nicht mehr vorab festgelegte Aufgaben vorzubereiten. Aufgaben in nicht vorher festgelegten Räumen können nun ausgeführt werden, etwa die Aufforderung, ein bestimmtes Objekt aus einem bestimmten Raum zu holen, wenn das Intelligenzsystem eine selbstgesteuerte Wahrnehmung der eigenen Position und Bewegung im Raum hat. Man spricht deshalb nicht mehr vom tiefen Lernen (deep learning), sondern vom langen Lernen (long learning). Gelernt wird längst nicht nur an Texten, sondern auch an Bilder-, Video- und Audiodaten. Multimodale Systeme in handelsüblichen Browsern wechseln bereits jetzt vom Text zum Bild und weiter zum Ton und wieder zurück. AGI scheint daher nicht mehr völlig unmöglich zu

sein. Die Entwicklung schreitet irritierend schnell voran und das nicht nur für die Öffentlichkeit.

3. Parallele Intelligenzen

‚Redteaming' nennt man in Technologiefirmen das Zusammenrufen der besten Köpfe aus unterschiedlichen Abteilungen, um die bedenklichen Folgen aktueller Entwicklungen abschätzen zu können. Solches Redteaming für die Folgenabschätzung scheint in den angesagten Techfirmen regelmäßig angesetzt zu werden. Auch wenn davon nur wenig nach außen dringt und Modelle wie ChatGPT-4 derzeit noch nicht veröffentlicht sind, ist die Richtung der Entwicklung in Umrissen klar und das schon vor der Veröffentlichung von ChatGPT-3 im November 2022, wenn auch unbeachtet von der Öffentlichkeit. Die neue Entwicklung ging schon vor dem Release von ChatGPT-3 in Richtung ‚Embodied AI', einer künstlichen Intelligenz, die sich durch Räume selbständig bewegt, Erfahrungen macht, Pläne schmiedet und so etwas wie einen Common Sense darüber erwirbt, wie unsere Welt funktioniert (LeCun, 2022; Sharma 2024). Diese autonome Maschinenintelligenz ist nicht die Weiterentwicklung der derzeit in der Öffentlichkeit diskutierten Großen Sprachmodelle und ihres Reinforcement Learnings, sondern plant auch unter unsicheren Randbedingungen ihr nächstes Verhalten. Wie ein solches System genauer aussieht, ist derzeit nur ansatzweise umrissen, ja wesentliche Voraussetzungen für eine solche Intelligenz fehlen, weil noch nicht einmal in Ansätzen klar ist, wie Common Sense computationell modelliert werden kann, also das Wissen um die Physik der Dinge, die sozialen Verhältnisse zwischen den Organismen, noch die intrinsischen Wertvorstellungen, die uns helfen, unsere Handlungen zu planen und auf unvorhergesehene Entwicklungen angemessen reagieren zu können.

Die Diskussion um gegenwärtige KI-Modelle führt also eher in die Irre, wenn man abschätzen will, wie weit KI-Systeme auch etwa für die Psychotherapie an Bedeutung gewinnen könnten. Nicht die Large Language Models sind zu diskutieren, stattdessen sind die grundsätzlichen Limitierungen künstlicher Intelligenz abzuwägen, wenn diese tatsächlich eine ‚Embodied AI' werden sollte (Ledezma et al., 2023). Das Investment von OpenAI in die norwegische Robotikfirma 1XTechnologies zielt in eben diese Richtung der Entwicklung einer Embodied AI (Sharma, 2024). Zu diesen fundamentalen Limitierungen der KI gehört

die Frage danach, wie die in der Evolution der Hominiden erworbenen und daher angeborenen Fähigkeiten des Menschen angefangen von der Orientierung im Raum bis zur intuitiven Wahrnehmung und Bewertung sozialer Beziehungen (Spelke, 2022) computationell modelliert werden könnte, denn Trainingsdaten aus der Evolution stehen nicht direkt zur Verfügung. Zweitens müsste KI trotz ihrer erheblichen Fehleranfälligkeit in die soziale Welt entlassen werden, um dort Erfahrungen ähnlich wie menschliche zu machen. Wir sehen schon bei der Frage des autonomen Fahrens, dass das nicht unbedingt eine gute Idee ist. Dennoch laufen Tests beispielsweise mit intelligenten Brillen gerade an, die visuelle und auditorische Daten in natürlichen Umgebungen integrieren (Meta, 2023). Drittens sind unsere Kenntnisse über den Aufbau der menschlichen Intelligenz noch so eingeschränkt und ihrerseits vielfach fehlerhaft oder unklar, dass die Modellierung der künstlichen in der menschlichen Intelligenz keine hinreichende Blaupause findet. Die hier genannten Limitierungen geben keine vollständige Liste möglicher grundsätzlicher Beschränkungen der KI, aber sie verdeutlichen, warum die Meinungen selbst unter Fachleuten divergieren, ob eine dem Menschen überlegene Artificial Superintelligence (ASI) überhaupt möglich ist bzw. wann dieser Punkt erreicht werden könnte, nie oder doch schon in fünf Jahren.

4. Am Ende Künstliche Psychotherapie?

Künstliche Intelligenz verfügt zumindest derzeit über kein Selbst und bildet kein Selbstverhältnis aus, wie es für jede psychotherapeutische Arbeit eine notwendige Bedingung ist. Trotzdem ist Künstliche Intelligenz für viele psychotherapeutische Anwendungen von Nutzen, angefangen von ihrer permanenten Verfügbarkeit bis zu einer unendlichen Geduld, mit der diese Systeme auf die Erwartungen der Patient*innen eingehen. Auch ihre diagnostischen Fähigkeiten nehmen schnell zu (Karthikesalingam & Natarajan, 2024). Entsprechende Anwendungen neben der Diagnostik wie das Erstellen von Interviews, der Dokumentation und das Notizen machen, der Unterstützungen von Patient*innen bei dem Ordnen der eigenen Gedanken und Gefühle für die nächste Sitzung, dann auch der Unterstützung der psychotherapeutischen Forschung versprechen, dass Künstliche Intelligenz bereits jetzt schon vielerorts selbstverständlicher Teil des psychotherapeutischen Prozesses ist (Tu et al., 2024). Sorgen, dass angesichts des riesigen Marktes für psychotherapeutische Anwendungen medizinische Standards und Prüf-

verfahren unterlaufen werden, liegen auf der Hand (Stade et al., 2023). Undeutlich ist gegenwärtig noch, wie trotz der Tendenz zu Fehlern und zur Trivialität, mit der KI stets den kleinsten gemeinsamen Nenner zum Maßstab nimmt, eine kreative Eigendynamik der KI-Systeme einzukalkulieren ist (Breithaupt et al., 2024; Meincke et al., 2024). Es liegt nahe, dass KI-Systeme eigene therapeutische Stile ausbilden werden, wenn sie erst einmal mit genug Daten aus therapeutischen Sitzungen gefüttert worden sind, – soweit muss man nicht viel spekulieren.

Doch setzen alle diese Überlegungen voraus, dass KI-Systeme in etwa entlang der bisherigen großen Sprach- und Bildmodelle konzipiert werden. Wahrscheinlicher als ein nächstes ChatGPT-5 oder 6 sind jedoch selbstexplorativ angelegte KI-Systeme. Wenn die hier skizzierten Überlegungen nicht ganz in die Irre gehen, dann ist schon in den nächsten fünf Jahren mit anderen KI-Formen zu rechnen, die so etwas wie ein ‚I'-Selbst und ein ‚Me'-Selbst entwickeln könnten. Ich nenne sie eine komplementäre Form der Intelligenz, denn so sehr sich diese KI-Systeme an die menschliche Intelligenz anlehnen, sie bleiben vielfach von dieser grundlegend verschieden und werden doch unsere Intelligenz ergänzen.

Im Feld der Medizin und der Psychotherapie im Besonderen ist schon jetzt zu erkennen, dass vor allem für Aufgaben der Diagnostik Systeme wie AMIE in der Entwicklung sind, die in realweltlichen Kontexten wie etwa Patient*innengesprächen eine reliable Diagnose erstellen können (Karthikesalingam & Natarajan, 2024). AMIE, das ist ein „Articulate Medical Intelligence Explorer", der für unterschiedliche medizinischen Situationen und deren Akteure skalieren soll. Eine spezifische Herausforderung für dieses System ist es dabei, vernünftige Schlussfolgerungen aus wechselnden Konversationsteilnehmer*innen zu ziehen, also den Sprecher*innen über ihre verschiedenen Rollen in einer Konversation hinweg zu folgen und den Wechsel der Sprecher*innen und ihren argumentativen Positionen und Perspektiven so zu folgen, dass daraus eine begründete Diagnose entsteht. Dass dann der Weg zur Therapie nicht mehr so weit ist, darf angenommen werden. Ob KI-Systeme dann auch die besseren Therapeut*innen werden, hängt wesentlich davon ab, ob Künstliche Intelligenz ein Selbstverhältnis ausbilden kann. Derzeit kann keine KI etwas Vergleichbares. Ob es hier eine prinzipielle Grenze gibt, ist umstritten. Komplementäre Intelligenz ist jedenfalls nicht gänzlich unmöglich. Nur wie diese Komplementarität auch in der Psychotherapie in fünf Jahren aussieht, ist Gegenstand vieler Spekulationen und nicht weniger wissenschaftlicher und technischer Anstrengungen.

Das alles muss irritieren, wenn nicht sogar verstören. Aber Redteaming ist nicht nur die Aufgabe von Techfirmen, sondern auch die nicht geringe Aufgabe der Psychotherapie. Sie hat die Expertise, um in diesem Prozess der Entwicklung einer komplementären Intelligenz nicht nur wie bisher Zuschauerin zu bleiben, sondern aktiv ihre Stimme zu erheben. Dieses Wissen um das Selbst und seinen nicht wenigen Störungen in die Entwicklung der komplementären Künstlichen Intelligenz einzubringen, ist nicht die zukünftige, sondern schon jetzt die gegenwärtige Herausforderung für die Psychotherapie und ihre Institutionen: Redteaming ist jetzt.

Referenzen

Abrams, Zara (2024). Monetizing Mental Health Technology. *APA, 55*(1), 76. https://www.apa.org/monitor/2024/01/trends-challenges-monetizing-mental-health (letzter Zugriff 29.04.2024).

Achiam, J. et al. (2023). GPT-4 Technical Report (v. 4). *arXiv*:2303.08774 [cs.CL]. https://doi.org/10.48550/arXiv.2303.08774 bzw. https://openai.com/research/gpt-4 (letzter Zugriff 29.04.2024).

Ananthaswamy, Anil (2024). New Theory Suggests Chatbot Can Understand Text. *Quanta Magazine.* https://www.quantamagazine.org/new-theory-suggests-chatbots-can-understand-text-20240122/ (letzter Zugriff 29.04.2024).

BabyLM (2023). BabyLM Challenge. Sample-Efficient Pretraining on a Developmental Plausible Corpus. https://babylm.github.io/ (letzter Zugriff 29.04.2024).

Bender, Emily et al. (2021). On the Danger of Stochastic Parrots. Can Language Models Be Too Big? *FAccT '21. Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. https://doi.rog/10.1145/3442188.3445922

Breithaupt, F. et al. (2024). Humans Create More Novelty than ChtGPT When Asked to Retell a Story. *Scientific Report, 14*(875). https://doi.org/10.1038/s41598-023-50229-7

Deutsche Gesellschaft für Psychologie (2023). Stellungnahme der Deutschen Gesellschaft für Psychologie auf die Schrift des Deutschen Ethikrates zu „Mensch und Maschine – Herausforderungen durch Künstliche Intelligenz". KI-basierte Systeme als Ersatz für Psychotherapie? Ein eindeutiges Nein! https://www.dgps.de/fileadmin/user_upload/PDF/Stellungnahmen/DGPs-Stellungnahme-Ethikrat_20232703.pdf (letzter Zugriff 29.04.2024).

Ebert, David, & Baumeister, Harald (2023). *Digitale Gesundheitsinterventionen: Anwendungen in Therapie und Prävention*. Springer.

Friston, Karl, & Frith, Christopher (2015). Active inference, communication and hermeneutics. *Cortex, 68*, 129–143. http://dx.doi.org/10.1016/j.cortex.2015.03.025

Javaheripi, Mojan, & Bubeck, Sébastian (2023). Phi-2: The Surprising Power of Small Language Models. *Microsoft Research Blog*. https://www.microsoft.com/en-us/research/blog/phi-2-the-surprising-power-of-small-language-models/ (letzter Zugriff 29.04.2024).

James, William (1890). *The Principles of Psychology*. New York: Henry Holt and Company.

Karthikesalingam, A., & Natarajan, V. (2024). AMIE: A Research AI System for Diagnostic Medical Reasoning and Conversations. Blog. *Google Research* (January 12). https://blog.research.google/2024/01/amie-research-ai-system-for-diagnostic_12.html?m=1 (letzter Zugriff 29.04.2024).

Knight, Will (2023). OpenAI's CEO Says the Age of Giant AI Models is Already Over. *Wired* (April 17). https://www.wired.com/story/openai-ceo-sam-altman-the-age-of-giant-ai-models-is-already-over/ (letzter Zugriff 29.04.2024).

LeCun, Yann (2022). A Path Towards Autonomous Machine Intelligence. *OpenReview* https://openreview.net/forum?id=BZ5a1r-kVsf (letzter Zugriff 29.04.2024).

Ledezma, F. D. et al. (2023). Machine Learning–Driven Self-Discovery of the Robot Body Morphology. *Science Robotics, 8*(85). https://doi.org/10.1126/scirobotics.adh0972

Lewis, M., & Mitchell, M. (2024). Using Counterfactual Tasks to Evaluate the Generality of Analogical Reasoning in Large Language Models. *arXiv*: 2402.08955 [cs.AI].

Mei, Qiouzhu, Xie, Yutong, Yuan, Walter, & Jackson, Matthew (2024). A Turing test of whether AI chatbots are behaviorally similar to humans. *PNAS, 121*(9), e2313925121.

Meincke, Lennart & Mollick, Ethan R., & Terwiesch, Christian (2024). Prompting Diverse Ideas: Increasing AI Idea Variance. *Social Science Research Network* (January 27). https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4708466 (letzter Zugriff 29.04.2024).

Meta (2023). Introducing the New Ray-Ban | Meta Smart Glasses. Blog. https://about.fb.com/news/2023/09/new-ray-ban-meta-smart-glasses/ (letzter Zugriff 29.04.2024).

Mishra, Chinmaya et al. (2023). Real-time Emotion Generation in Human-Robot Dialogue Using Large Language Models. *Frontiers in Robotics and AI,* 10. https://doi.org/10.3389/frobt.2023.1271610

Ornes, Stephen (2023). The Unpredictable Abilities Emerging From Large AI Models. *Quanta Magazine.* https://www.quantamagazine.org/the-unpredictable-abilities-emerging-from-large-ai-models-20230316/ (letzter Zugriff 29.04.2024).

Pavlus, John (2019). Machines Beat Humans on Reading Tests. But Do They Understand? *Quanta Magazine*. https://www.quantamagazine.org/machines-beat-humans-on-a-reading-test-but-do-they-understand-20191017/ (letzter Zugriff 29.04.2024).

Shaeffer, Ryan, Miranda, Brando, & Koyejo, Sanmi (2023). Are Emergent Qualities of Large Language Models a Mirage? *arXiv*, arXiv:2304.15004 [cs.AI]. https://doi.org/10.48550/arXiv.2304.15004

Sharma, Shubham (2024). 1X, Robotic Startup Backed by OpenAI, receives $ 100M in Funding. *VentureBeat* (January 11). https://venturebeat.com/ai/1x-robotic-startup-backed-by-openai-receives-100m-in-funding/ (letzter Zugriff 29.04.2024).

Spelke, Elizabeth (2022). *What Babies Know. Core Knowledge and Composition*. Oxford University Press.

Stade, Elizabeth et al. (2023). Large Language Models Could Change the Future of Behavioral Healthcare: A Proposal for Responsible Development and Evaluation. *PsyArXiv*. https://doi.org/10.31234/osf.io/cuzvr

Tu, Tao et al. (2024). Towards a Conversational Diagnostic AI. *arXiv*:2401.05654 [cs.AI]. https://doi.org/10.48550/arXiv.2401.05654

Villalobos, Pablo et al. (2023). Will We Run Out of Data? An Analysis of the Limits of Scaling Datasets in Machine Learning. arXiv:2211.04325 [cs.LG]. https://doi.org/10.48550/arXiv.2211.04325

Wolfram Alpha (2023). What is ChatGPT Doing … And Why Does It Work? Blog. https://writings.stephenwolfram.com/2023/02/what-is-chatgpt-doing-and-why-does-it-work/ (letzter Zugriff 29.04.2024).

Wei, Jason (2023). 137 Emergent Abilities of Large Language Models. Blog. https://www.jasonwei.net/blog/emergence (letzter Zugriff 29.04.2024).

Yu, Dingli et al. (2023). Skill Mix. A Flexible and Expandable Family of Evaluations for AI Models. *Open Review.* https://openreview.net/forum?id=Z05m9cRpRa&referrer=%5Bthe%20profile%20of%20Anirudh%20Goyal%5D(%2Fprofile%3Fid%3D~Anirudh_Goyal1) (letzter Zugriff 29.04.2024).