Wie intelligent kann Künstliche Intelligenz sein?

Klaus Niedermair[1]

[1] Leopold-Franzens-Universität Innsbruck

Zusammenfassung

Ohne Zweifel bringt Künstliche Intelligenz (KI) mittlerweile Leistungen, die menschliche Intelligenz vielfach unterstützen, kompensieren und zum Teil ersetzen können. Dieser Beitrag stellt die Frage, wo die prinzipiellen Grenzen der KI liegen und ob überhaupt von Intelligenz im eigentlichen Wortsinn die Rede sein kann. Wenn menschliche und künstliche Intelligenz nur anhand ihrer Symptomatik beurteilt werden, nivelliert sich zusehends der Unterschied bzw. schlägt das Pendel schon in die andere Richtung aus. Hier ist wohl ein reduzierter Begriff von Intelligenz im Spiel. Anhand der Hermeneutik und der Tractatus-Philosophie von Ludwig Wittgenstein wird gezeigt, dass ein KI-System wie ChatGPT grundsätzliche Defizite aufweist. Es kann Sprache nicht verstehen und nicht lernen. Schlüsse ziehen kann es zwar deduktiv als Expertensystem, jedoch induktiv nur statistisch und abduktiv gar nicht. Zudem ist ein KI-System wie ChatGPT ein geschlossenes selbstreferenzielles System, das zu Paradoxien und Selbstreplikation durch Autophagie tendiert, konservativ und anfällig für Manipulation ist und à la longue Qualität und Reichweite unserer Wirklichkeitserkenntnis reduziert. Gefragt sind ein technologiekritischer Ansatz, die Überwindung des behavioristischen und naiv technizistischen Paradigmas der KI sowie eine kritische Metaphernanalyse der Anthropomorphismen der KI. So können Grundlagen geschaffen werden für eine sinnvolle Sprache über KI und damit auch für brauchbare theoretische Konzepte und Modelle ihrer Nutzung und Weiterentwicklung.

Niedermair, Klaus. (2026). Wie intelligent kann Künstliche Intelligenz sein? In Klaus Rummler, Günther Pallaver, Petra Missomelius, Valentin Dander, & Oliver Leistert (Hrsg.), *Streifzüge an den Nahtstellen von Medien, Bildung und Philosophie* (2. erw. Aufl., Medien – Wissen – Bildung 2025, S. 39–74). Zürich: OAPublishing Collective. https://doi.org/10.21240/978-3-03978-164-5_05

How Intelligent can Artificial Intelligence be?

Abstract

Undoubtedly, Artificial Intelligence (AI) now achieves feats that can extensively support, compensate for, and partially replace human intelligence. This paper raises the question of where the principal limits of AI lie and whether it can even be spoken of as „intelligence“ in the true sense of the word. If human and artificial intelligence are judged solely by their symptoms, the distinction increasingly blurs, or the pendulum already swings in the other direction. This likely points to a reduced concept of intelligence at play. Drawing on hermeneutics and Ludwig Wittgenstein's Tractatus-philosophy, it will be shown that an AI system like ChatGPT exhibits fundamental deficits. It cannot understand language and cannot truly learn. While it can perform deductive reasoning as an expert system, its inductive reasoning is merely statistical, and it lacks abductive reasoning entirely. Furthermore, an AI system like ChatGPT is a closed, self-referential system that tends towards paradoxes and self-replication through autophagia; it is conservative and susceptible to manipulation, and over time, it reduces the quality and scope of our understanding of reality. What is needed is a technology-critical approach, an overcoming of the behaviorist and naively technocentric paradigm of AI, and a critical analysis of the anthropomorphic metaphors used for AI. This can lay the groundwork for a meaningful language about AI and, consequently, for useful theoretical concepts and models for its utilization and further development.

Wie intelligent kann Künstliche Intelligenz sein?

Spätestens seit der Veröffentlichung von *ChatGPT* Ende 2022 ist Künstliche Intelligenz (KI) omnipräsent. Doch unsere Lebenswelt ist schon länger und zunehmend mit KI infiltriert und von KI abhängig, von der alltäglichen KI im Smartphone bis zur KI in der medizinischen Diagnostik, in der industriellen Fertigung und Produktion, in der Textgenerierung, in der Videoüberwachung usw. Im öffentlichen Diskurs, in den Medien, in wissenschaftlichen Publikationen ist immer wieder die Rede von der rasanten Entwicklung der KI, von den großen Veränderungen für die Gesellschaft und für den Einzelnen, von den „Chancen“, „Herausforderungen“, „Risiken“ – und plakativ wird „kritische Reflexion“ gefordert. In solchen Stehsätzen zeigt sich

auch ein Stück weit die Überforderung und die Hilflosigkeit angesichts einer Entwicklung, die chaotisch und disruptiv verläuft, auch weil sie vorwiegend profit- und marktgetrieben ist, dementsprechend disparat und kontroversiell sind die Interessen, Ziele und Projekte rund um die KI.

Es gibt rundum einen KI-Boom, eine Aufbruchsstimmung, Hoffnungen, Erwartungen, Utopien, Dystopien, Ängste, Untergangsszenarien etc. Regierungen, Unternehmen, Universitäten usw. wollen mit im Spiel sein, KI wird zum Schlüsselfaktor in der Wirtschaft, in der Bildung. Doch die KI wird auch als undurchschaubare Blackbox wahrgenommen; nicht viele wissen, wie sie funktioniert, niemand weiß, wohin sie sich weiterentwickeln wird. Es gibt vielerlei Unsicherheiten, Bedenken und Ängste, z. B. im Hinblick auf Datenschutz, Überwachung, Kontrolle, Manipulation, Desinformation, Deepfakes, Propaganda usw. Zudem tendiert der KI-Markt eindeutig in Richtung Privatisierung, Deregulierung und Volatilität, forciert von der neuen US-Regierung. Und das Pünktchen auf das „i" der KI setzt der Hightech-Milliardär und Hobby-Politiker Elon Musk, der verkündet, dass die sogenannte *Artificial General Intelligence* (AGI), eine Steigerung der AI, bereits 2025 intelligenter sein wird als der intelligenteste Mensch, das hieße dann auf jeden Fall wohl auch: intelligenter als die Protagonisten der derzeitigen US-Administration.

Integrative Perspektiven sind gefragt, Ganzheitlichkeit, Inter- und Transdisziplinarität, kein Denken in Silos. Ein Schritt in diese Richtung wäre es schon, rein auf begrifflicher Ebene deutlicher zu unterscheiden zwischen der *Effizienz* und der *Effektivität* der KI. Daraus ergeben sich zwei unterschiedliche Fragen. Einerseits: Werden mit KI vorgegebene Ziele schnell und ressourcensparend erreicht? Andererseits: Sind die Ziele, die mit KI erreicht werden, sinnvoll für die Gesellschaft und den Einzelnen? Das sind dann verschiedene Diskurstypen. Im Hinblick auf technische Effizienz ist die KI zweifellos gut und wird immer besser, das beweist eindeutig die Symptomatik der Intelligenz. Es gibt Szenarien, in denen die Künstliche Intelligenz der menschlichen überlegen ist, bspw. klingen manche Texte von *ChatGPT* zweifellos besser als von Menschen geschriebene. Doch man kann fragen, ob in diesem Fall die KI, was ihre Effektivität

betrifft, um es einfach zu sagen, auch gut ist. Das Paradebeispiel: KI kann Leben retten in der Medizin durch die schnelle Auswertung von riesigen Datenpools, KI kann Leben zerstören durch einen taktisch gezielten Militäreinsatz. Beide Male steht die effiziente Nutzung der KI außer Frage, der wunde Punkt ist die Effektivität. Oder: Ist es sinnvoll, sich von *ChatGPT* Texte schreiben zu lassen? Ja und nein. So sind ausgeklügelte Techniken des *Prompt Engineering* zweifellos wichtig, wenn man *ChatGPT effizient* nutzen will, und es ist insofern geraten, sich entsprechende technische Medienkompetenz anzueignen. Aber man könnte fragen, ob es überhaupt *effektiv* ist, dass man mit *ChatGPT* in diese „Kommunikation" tritt, bspw. um nach Informationen zu suchen oder sich Texte zusammenfassen und schreiben zu lassen usw. Und ob man sich dabei nicht selbst in eine Komfortzone manövriert, zentrale Bereiche von Intelligenz auslagert und so, und das ist der Punkt, aus Faulheit und in Freiwilligkeit in diese Abhängigkeit und Fremdbestimmtheit geht und Selbstbestimmtheit aufgibt. Wer denkt da nicht an Immanuel Kant:

> „Unmündigkeit ist das Unvermögen, sich seines Verstandes ohne Leitung eines anderen zu bedienen. Selbstverschuldet ist diese Unmündigkeit, wenn die Ursache derselben nicht am Mangel des Verstandes, sondern der Entschließung und des Mutes liegt, sich seiner ohne Leitung eines anderen zu bedienen. Sapere aude!"

KI ist in den Medien zum Dauerbrenner geworden. Ein schneller Eindruck (er wäre natürlich erst inhaltsanalytisch zu belegen): Die Diskussion ist einseitig fokussiert, die Perspektive der Wirtschaft dominiert, KI-Unternehmen sind der Hit, Nachholbedarf wird beschworen, Thema ist bspw. auch die Veränderung in der Arbeitswelt – tiefergehende Themen wie Nachhaltigkeit, globale soziale Gerechtigkeit usw. sind unterrepräsentiert. Doch *en passant* gibt es in der Öffentlichkeit immer schon mehr Nachfrage nach einem Diskurs zu Fragen *jenseits* der *Effizienz* von KI, *jenseits* der vorwiegend markt- und betriebswirtschaftlich motivierten Implementierung von KI, *jenseits* der technischen Details, How-to-do's und individuellen Schlauheiten in der Nutzung der KI. Also Fragen zur *Effektivität* der KI und ihrer Anwendungsszenarien. Was sind in der Nutzung von KI bspw. sinnvolle Standards von Medienkompetenz, sinnvolle

Normen in Ethik, Regulierung und Legislatur? Wie soll KI-Kompetenz (AI Literacy) konkret ausformuliert werden? Eine effiziente und technisch kompetente Nutzung von KI ist wichtig. Inwieweit ist aber auch die effektive, sinnvolle Nutzung von KI im Fokus? Wie werden die erforderlichen Kompetenzen konkretisiert, wenn man über die (leider oft nur plakativen) Formulierungen wie „Reflexion", „kritischer Auseinandersetzung" usw. hinausgehen will? In diesem Zusammenhang gibt es einige Desiderata (Hug 2024, S. 141–142): Überwindung der Werkzeugperspektive und der primär lerntechnologischen Didaktiken, Berücksichtigung medienanthropologischer Dimensionen (KI-Nutzung prägt die Wirklichkeit), Beachtung tiefenpsychologischer Dimensionen. Des Weiteren: Wie funktioniert Influencing und Meinungsbildung zum Thema KI in der Öffentlichkeit? Wie soll in Medien und in Bildungsinstitutionen das Thema KI präsentiert und diskutiert werden? Vorwiegend affirmativ und technikzentriert oder mehr kritisch im Hinblick auf mögliche Folgen der KI für die Gesellschaft und den Einzelnen? Welches Experten- und Orientierungswissen kann den Entscheidungsträgern in Politik und Wirtschaft angeboten werden? Konsens dürfte wohl sein: Die Wissenschaften sollten von der Politik ein klares Mandat bekommen, um objektiv, demokratisch und „interesselos" theoretische Grundlagenarbeit machen und in allen Fragen der Entwicklung und Nutzung von KI mitgestalten zu können. Nur Markt und Privatisierung ist toxisch, nicht allein die Oligarchen Musk, Bezos, Zuckerberg und Konsorten dürfen die Weichen stellen.

Dieser Beitrag möchte einen Versuch in die Richtung einer integrativen Perspektive wagen. Im Mittelpunkt steht die immer noch aktuelle Frage: Was ist der Unterschied zwischen der menschlichen Intelligenz und der sogenannten künstlichen Intelligenz? Oder ist vielleicht eine solche Frage ohnedies unsinnig? Im Detail geht es darum, was menschliche Intelligenz *ist* und was Künstliche *sein könnte* oder was Künstliche Intelligenz *ist* und was die menschliche *noch sein kann*. Was als menschliche Intelligenz definiert wird, ist natürlich entscheidend dafür, *ob überhaupt* und wenn ja, *wie sehr* ein KI-System, intelligent sein kann. Die Frage *ob überhaupt* ist bereits im Titel dieses Beitrages beantwortet: „Wie intelligent kann Künstliche Intelligenz sein?" konzediert, dass KI bereits Bereiche von Intelligenz abdeckt.

Die Frage ist nur mehr *wie sehr*, und vor allem, was übrigbleibt als exklusiv menschliche Intelligenz. Noch nie in der Geschichte der Menschheit ist es dem Menschen in dieser Hinsicht – in seinem Selbstverständnis, in seiner Identität, und dazu gehört ja auch die Intelligenz – so ans Leder gegangen, noch nie waren die Wissenschaften so unter Druck, ein (neues) Menschenbild zu (er)finden. Die Philosophen haben sich (ohne diesen Druck) immer schon mit Erkenntnis, Sprache, Wissen, Wissenschaft, Intelligenz usw. beschäftigt – Aristoteles, Platon, Descartes, Leibniz, Kant, Wittgenstein, um nur ein paar zu nennen – und uns ein reiches theoretisches Potenzial in die Hand gegeben, wie wir angesichts der KI weiterdenken könnten. In diesem Sinne will ich auf der Grundlage meiner Interpretation des Tractatus von Ludwig Wittgenstein (1921) ein paar Gedanken zur Intelligenz der KI formulieren und einige dann mit *ChatGPT* sozusagen im Hybrid-Modus – auf der Suche nach Synergien zwischen menschlicher und künstlicher Intelligenz – diskutieren. Der Fokus ist sprachphilosophisch, die Sprache ist das Medium der Intelligenz, wie auch Liang Wenfeng, CEO und Mastermind von *DeepSearch*, dem neuesten Hit am KI-Markt, in einem Interview feststellt:

> „... we believe that human intelligence might fundamentally be about language – that human thought might simply be a process of language. What you perceive as thinking might just be your brain weaving language."[1]

Wie ernst ist der Begriff „Künstliche Intelligenz" eigentlich gemeint? Dieser Terminus technicus hat sich seit den 1960er-Jahren etabliert, ursprünglich in der Welt der Informatik und Mathematik, mithin ein autochthon technischer Begriff, auch in diesem Sinn technicus. Die Geburtsstunde der Künstlichen Intelligenz ist die sog. *Dartmouth Conference*. Die KI-Pioniere John McCarthy, Marvin Minsky, Nathaniel Rochester und Claude Shannon waren die Organisatoren dieser Veranstaltung, ihr *Porposal* für die Beantragung einer Förderung bei der *Rockefeller Foundation* beginnt mit dem Satz:

> „We propose that a 2 month, 10 man study of artificial intelligence be carried out during the summer of 1956 at Dartmouth

1 https://jiexu.substack.com/p/interview-deepseek-founder-liang-wenfeng-a-journey-by-curiosity.

> College in Hanover, New Hampshire. The study is to proceed on the basis of the conjecture that every 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." (McCarthy u. a. 1955)

Schon in der ersten Aufbruchsstimmung hat man „Künstliche Intelligenz" wohl auch wörtlich so gemeint. Doch auch wenn die Maschine Merkmale von Intelligenz simulieren kann, ist und bleibt der Begriff „Künstliche Intelligenz" eine Metapher. Die Basis ist eine einfache Analogiebeziehung: Die Intelligenz verhält sich zum Menschen, wie die IT-Leistung zur IT-Maschine; aufgrund dieser Analogie wird dann der IT-Maschine Intelligenz attestiert und mit der Etikette „Künstliche Intelligenz" versehen. Also eine metaphorische Übertragung ganz dem Beispiel entsprechend, das Aristoteles gibt:

> „Das Alter verhält sich zum Leben, wie der Abend zum Tag; der Dichter nennt also den Abend ‚Alter des Tages', oder das Alter ‚Abend des Lebens'". (Niedermair 2001, S. 144)

Wie geht man mit dieser Metapher um? Man kann sie wörtlich nehmen und so die der Künstlichen Intelligenz als Realität zementieren. Oder man kann die metaphorische, anthropomorphe Sprache, die prägend ist für das Bild der KI in der Öffentlichkeit, mithilfe einer *Metaphernanalyse* (Niedermair 2001) ins Visier nehmen und Mythologien, Hypes und Missverständnisse, die in den letzten zwei Jahren parallel mit der rasanten Entwicklung der KI zugenommen haben, *technologiekritisch* (Niedermair 1998) reflektieren und ggf. versuchen, als Alternative eine objektive und klare Sprache über KI zu entwerfen. Metaphern sind eine Bereicherung der Sprache, sie erweitern unsere Möglichkeiten der Erkenntnis, sie können auch in die Irre führen, es gibt problematische Metaphern, „Künstliche Intelligenz" ist wohl auch eine. Vermutlich hat sich diese Metapher etabliert, um der KI-Maschine einen schönen Namen zu geben, sie quasi gesellschaftsfähig zu machen, als vertraut zu präsentieren und zu verschleiern, dass die KI eigentlich auch eine mysteriöse Blackbox ist. Es ist schwer, diese metaphorische Begrifflichkeit aus der Welt zu schaffen. Insofern wäre eine Sprachregelung sinnvoll: „Künstliche Intelligenz" wird als Terminus technicus verwendet für Systeme der

maschinellen Informationsverarbeitung, die Merkmale von Intelligenz – wie Lernen, Schlussfolgern oder Sprachverstehen – nachbilden und simulieren." Aber immer dann, wenn dieser Begriff wörtlich und nicht nur metaphorisch gemeint sein könnte, was z. B. bei Kleinschreibung naheliegend ist, soll hier von der *„sog.* künstlichen" Intelligenz die Rede sein – aus Gründen der Begriffshygiene, denn:

> „KI entscheidet, lernt, handelt, denkt etc. trotz entsprechend semantischer Suggestionen genauso wenig, wie ein autonomes Auto autonom ist, wie ein Roboter zu etwas gezwungen werden könnte oder wie Informatiker, KI-Forscher und Data Scientists mit dem Schaffen von verblüffenden IT-Systemen zu Schöpfergöttern würden." (Gransche & Manzeschke 2024, S. 76)

Trotzdem: Was ist tatsächlich der Unterschied zwischen der menschlichen und der sog. künstlichen Intelligenz? Woran könnte man einen solchen überhaupt festmachen? Wenn es um Intelligenz geht, sehen wir ja immer nur die Oberfläche, also bestenfalls ein intelligentes Verhalten, eine Symptomatik von Intelligenz, die man anhand bestimmter Kriterien messen kann, egal ob menschliche oder sog. künstliche. Und wenn *beide* gleiche Symptome zeigen, und das ist zweifellos immer öfter der Fall, unterscheiden sie sich dann wirklich noch und wenn ja worin? Diese Frage war auch der Fokus einer Forschungsarbeit von Alan Turing (1950). Damals hatte diese Frage zwar nicht die Aktualität wie heute, intelligente Systeme gab es noch nicht wirklich, auch nicht den Terminus technicus „Künstliche Intelligenz". Trotzdem gab es in der Theorie die Frage, ob künstliche Intelligenz möglich ist. Turing erfand ein Test-Szenario für den Nachweis von Intelligenz. Mehrere Personen führen den Test durch. Jede Person interagiert mit zwei Spielern A und B, und zwar nur schriftlich und ohne sie zu sehen, wobei einer der beiden Spieler eine Maschine ist und der andere ein Mensch. Wenn nun mindestens 30 % der Personen, die den Test durchführen, davon überzeugt sind, dass sie mit einem Menschen sprechen, obwohl es die Maschine war, dann hat die Maschine den Turing-Test bestanden. Oder mit anderen Worten: Wenn beide, A und B, in ihren Antworten Symptome von Intelligenz zeigen und keine wesentlichen Unterschiede in ihrem intelligenten Verhalten erkennbar sind, dann hat auch die Maschine den „Intelligenztest" bestanden, d. h. sie verfügt über Intelligenz. Turing selbst setzt dies

ohnedies als Selbstverständlichkeit voraus, sodass für ihn eigentlich schon die Frage obsolet ist, ob eine Maschine Intelligenz haben könne.

Natürlich hätte dieser Test nur dann Sinn, wenn ein wesentlich reduzierter Begriff von Intelligenz vorausgesetzt wird, nämlich dass äußerlich erkennbare Symptome eines intelligenten Verhaltens ausreichen, um Intelligenz nachzuweisen. Die Absurdität einer solchen Reduktion von Intelligenz und mithin auch dieses „Intelligenztestes" zeigt sich schon allein darin, dass mittlerweile die KI so weit fortgeschritten ist, dass alles in die andere Richtung gekippt ist. Es ginge mittlerweile in einem Turing-Test nicht mehr darum, ob bewiesen werden kann, dass die KI *mindestens so viel* Intelligenz hat wie ein Mensch, sondern ob sie vortäuschen kann, dass sie *nicht mehr* Intelligenz hat als ein Mensch. Das heißt die KI muss dann quasi so *meta*-intelligent sein, ihre Symptome von Intelligenz so steuern zu können, dass sie nicht aufgrund ihrer Superintelligenz als künstliche auffliegt; sie muss also ihre Intelligenz ein paar Grade herunterdrehen; quasi auf menschliches Niveau, also z. B. die Komplexität ihres Outputs reduzieren und in die stilistische Glattheit (negativ formuliert: in die Fadesse) ihrer Diktion ein paar menschliche Ecken und Kanten einbauen usw. Noch ein Sarkasmus *marginaliter* sei erlaubt: Es ist davon auszugehen, dass in Hinkunft studentische (und vermutlich nicht nur solche, sondern auch andere) wissenschaftliche Arbeiten zunehmend nicht nur mit KI-Textbausteinen in kleinen Dosen garniert, sondern in ihrer Substanz überhaupt mit KI-Texten in größerem Ausmaß realisiert werden, und dass insofern die Nachfrage nach Tipps und Ratgebern für Prompts und auch brauchbare Verschleierungsstrategien steigt, die ggf. wiederum künstlich intelligent generiert werden usw. Aber um es zu beenden, klar ist ja: Simulieren ist nicht gut; besser ist, den eigenen Stil in Denken und Sprache finden, im lebensweltlichen, politischen, persönlichen Kontext reflektieren, Authentizität und Kreativität genießen.

Allerdings: Lässt sich überhaupt das Argument halten bzw. wie lange, dass KI „nur" eine perfekte Symptomatik von Intelligenz produziert, also Intelligenz simuliert, aber nicht „wirklich" intelligent ist? Was kann das ominöse Residuum menschlicher Intelligenz sein? In vielen Bereichen gibt es das nicht mehr, Rückzugsgefechte scheitern kläglich. Was will man bspw. von einem Rechtsanwalt mehr, wenn

man mindestens die gleiche, wenn nicht die bessere Rechtsauskunft und Expertise schnell, bequem, kostengünstig auch von einer KI haben kann, die mit einem so einen großen Korpus an juristisch relevanten Texten operiert, das der Rechtsanwalt vermutlich so schnell nicht im Blick haben dürfte. Ist es ein Mehrwert des menschlichen Rechtsanwalts, wenn er quasi Beziehungsarbeit anbietet, Empathie, soziale Kompetenz, Verhandlungsgeschick, Charisma? In der Regel ist das in einer juristischen Dienstleistung nicht gefragt. Und was nützt da die große Theorie, könnte man fragen, die auf den Unterschied menschlich und künstlich beharrt, z. B. mit dem Hinweis, dass die KI – wenn man sie sprachwissenschaftlich z. B. mit Noam Chomsky betrachtet – eigentlich nur eine syntaktisch halbwegs valide Sprachfassade inszeniert (auch die nur halbwegs, weil sie nicht mit einem Regelwissen à la Generativer Grammatik arbeitet, sondern nur mit Wahrscheinlichkeiten von Mustern, Vektoren, Layers à la Generativer KI), dass es keine Semantik gibt (der KI ist es egal, was die Wörter und Sätze bedeuten) und keine Pragmatik (der KI ist es egal, in welchem Kontext die Sprache verwendet wird). Ja, alle diese Bedenken haben theoretisch sehr großes Gewicht und dürfen nicht unterschlagen werden. Doch was zählt es, wenn der Mensch mit der KI zufrieden ist, wenn der Mensch in Wirklichkeit das, was der KI fehlt, ohne es zu wissen sogar proaktiv kompensiert, also das semantische Vakuum füllt, indem er dem KI-Output Bedeutung verleiht, und auch das pragmatische Vakuum füllt, indem er dankbar den KI-Output Relevanz und Valenz in seinen Interessens- und Gebrauchsszenarien zuerkennt, und – zu guter Letzt – wenn die Qualität der Performance stimmt (und das ist weitgehend der Fall), also Qualität, Sinn und Hype der KI-Nutzung noch glorreich bestätigt wird? Wenn der KI all das – semantisch die Bedeutung, pragmatisch der Sinn – egal ist, heißt das doch auch: Wie tiefgreifend auch immer die Auswirkungen der KI sein mögen (großflächig positiv, aber punktuell auch als möglicher Untergang der Menschheit), nicht vergessen werden darf: Die menschliche Intelligenz entwickelt die technische Maschine der sog. künstlichen *und* haucht dieser Maschine durch Zuschreibung, Projektion und Übertragung (tiefenpsychologisch) Leben ein. Die menschliche Intelligenz kompensiert durchaus kreativ die prinzipiellen Defizite der künstlichen (betreffend Sinnvakuum

usw.), indem sie sie als Intelligenz, wenn auch aus künstliche, benutzt. Warum tut man das? Aber zurück zum Turing-Test.

Die KI kann nicht Sinn verstehen oder Sinn meinen

Auch John Searle (1980) war mit dem Turing-Test nicht einverstanden. Die quasi-intelligente Maschine *versteht* nichts von dem, was sie rechnet und an quasi-intelligenten Symptomen produziert. Searle schlägt als alternatives Gedankenexperiment sein chinesisches Zimmer vor. Eine Person, die nicht Chinesisch kann, ist in einem Zimmer eingesperrt. Außerhalb des Zimmers befindet sich eine zweite Person, die Zettel mit chinesischen Fragen durch einen Briefschlitz in das Zimmer werfen kann. Die Person im Zimmer versteht zwar nicht, was auf den Zetteln steht, kann aber diesen Schriftzeichen anhand einer Tabelle Schriftzeichen mit einer chinesischen Antwort zuordnen und diesen Zettel nach außen reichen. Auch wenn die Person außerhalb des Zimmers glaubt, im Zimmer befinde sich ein Mensch, der chinesisch antwortet, beweist sein Verhalten noch lange nicht, dass er wirklich Chinesisch versteht. Das heißt auch wenn sich eine Maschine intelligent verhält und den Turing-Test besteht, heißt das noch nicht, dass sie etwas *versteht*. Sie zeigt bestenfalls nur eine „schwache", aber keine „starke Intelligenz". Searle hat also die Frage umgedreht. Es geht ihm nicht wie beim Turing-Test darum zu beweisen, dass ein KI-System über Intelligenz wie ein Mensch verfügt nur bloß aufgrund seines quasi-intelligenten Verhaltens. Searle geht von der Frage aus, ob es überhaupt notwendig ist, dass die Person im Zimmer Chinesisch kann (also in diesem Sinne intelligent ist), und arrangiert ein Setting, in dem die Person tatsächlich wie ein KI-System mit „schwacher Intelligenz" agiert, was für die Person außerhalb aber keinen Unterschied macht. Aber um „starke Intelligenz" geht es nicht.

Erkenntnis- bzw. wissenschaftstheoretisch bedeutet das: Im Sinne des Behaviorismus wäre laut Turing-Test eine Maschine, die sich so verhält, als wäre sie intelligent, und insofern Symptome von Intelligenz zeigt, tatsächlich intelligent. Im Sinne der Hermeneutik wäre sie – wenn man diese Unterscheidung macht – höchstens „schwach", nicht aber „stark" intelligent, denn das Auftreten dieser Symptome ist kein Beweis dafür, dass die Maschine versteht, was ihr gesagt wird, und meint, was sie sagt.

Damit scheint auch das KI-System *ChatGPT* d'accord zu sein, wie die folgende KI-Interaktion[2] zeigt. Ich verwende übrigens bewusst den Ausdruck „Interaktion", um zu betonen, dass es sich nicht um eine Kommunikation zwischen zwei Menschen handelt, obwohl ich selbst das im *ChatGPT* übliche Kommunikationssetting mitspiele und dem System, als sei es eine Person, diese Frage stelle:

⇒ „Ich habe eine erkenntnis- bzw. wissenschaftstheoretische Frage. Etwas erkennen, verstehen, meinen, lernen und tun erfordert Intelligenz. Bist du intelligent? Verstehst du auch, was ich schreibe, und meinst du auch, was du ‚schreibst'? Oder simulierst du mit dem, was du schreibst, Intelligenz?"

⇐ „Ich bin ein KI-Modell, das auf riesigen Mengen von Textdaten trainiert wurde. Meine Antworten basieren auf statistischen Mustern, die ich in diesen Daten gelernt habe. Das bedeutet, dass ich keine bewussten Gedanken, kein Selbstverständnis und kein Verständnis dessen habe, was ich schreibe. Ich generiere Text nur auf Grundlage von Wahrscheinlichkeiten und Mustern, die in den Daten vorhanden sind.

Ich simuliere Intelligenz durch den Einsatz von Algorithmen, aber ich verstehe nicht wirklich, was ich schreibe. Die Unterscheidung zwischen menschlicher Intelligenz und meiner ‚Intelligenz' ist genau das, was die wissenschaftstheoretische Diskussion um künstliche Intelligenz so faszinierend macht. Es stellt sich die Frage, was es wirklich bedeutet, zu verstehen, zu wissen und zu denken."

Kann es sein, dass ich *ChatGPT* mit meiner doch etwas suggestiven Frage bereits auf den richtigen Kurs gebracht habe? Zumindest beweist das künstlich intelligente Statement, dass es eine gute Simulation von Intelligenz ist. Allerdings irritiert unter anderem noch die Formulierung „Ich bin ...". Auch wenn dieses „Ich" wie gesagt zum üblichen KI-Kommunikationssetting gehört, ist eine solche Sprache nicht nur beiläufiges und kosmetisches Detail. Der Anthropomorphismus schafft eine irreführende Wirklichkeit, die einer sachorientierten, technischen, rationalen, verantwortungsbewussten, sinnvollen Nutzung von KI im Wege stehen kann. Es ist anzunehmen, dass die

2 OpenAI's ChatGPT AI language model, persönliche KI-Interaktion, 01.11.2024. Fragen sind hier mit ⇒ gekennzeichnet, Antworten mit ⇐.

Inszenierung einer informellen Kommunikation quasi zwischen zwei Personen eine Authentizität suggeriert, die dazu führt, dass Informationen eher glaubwürdig erscheinen und weniger kritisch hinterfragt werden – im Unterschied zu Printmedien bzw. Online-Medien mit statischen Inhalten, die distanzierter und kritischer rezipiert werden und insofern eine wissenschaftliche Herangehensweise fördern.

Doch sehen wir uns die prägnante „Selbsterkenntnis" von *ChatGPT* genauer an, indem wir erkenntnis- und wissenschaftstheoretisch überlegen, „was es wirklich bedeutet, zu verstehen, zu wissen und zu denken". Verstehen heißt, die Bedeutung bzw. den Sinn eines Zeichens erfassen. Ein Zeichen sein – und insofern Bedeutung bzw. Sinn haben – kann im Prinzip alles, sofern es als Zeichen *gebraucht* wird (und insofern auch ein Medium ist), z. B. eine sprachliche Artikulation (wir meinen und verstehen den Sinn von Lauten), ein Verhalten (wenn wir uns die Hand reichen, meinen und verstehen wir das als Begrüßung oder Bekräftigung oder Verabschiedung), irgendein mediales Artefakt (wir meinen etwas mit einem Text in einem Buch und wir verstehen Schrift usw.). Bedeutung und Sinn ist das, was jemand *meint*, wenn er ein Zeichen verwendet, und das wir dann – wenn alles gut geht – auch so, in dieser Bedeutung, *verstehen*. Wenn alles gut geht, denn so sicher ist das nicht: Zeichen sind eigentlich nur Symptome von irgendwas, d. h. wir haben grundsätzlich nicht die Sicherheit, dass wir diese Symptome, diese Zeichen, ein Verhalten, sprachliche Laute, Gesten usw. richtig verstehen – oder sie so meinen, dass eine Chance besteht, dass eine andere Person sie richtig, und das heißt so, wie wir sie meinen, versteht.

Meinen und Verstehen von Zeichen ist trotzdem möglich – so die Position der Hermeneutik – aufgrund eines *Vorverstehens*, eines *Vorverständnisses*, welches wir, wenn wir etwas verstehen, mit jemandem, der etwas meint, gemeinsam haben, da wir eine gemeinsame Lebensform, Kultur, Sozialisation haben, nicht im Ganzen gemeinsam, aber so viel, dass Kommunikation möglich ist. Darin besteht der hermeneutische Zirkel: Wir müssen ein Vorverständnis haben, also schon *etwas* vorher verstehen bzw. verstanden haben, bevor wir *etwas anderes* verstehen können. Dieser Zirkel zwischen Vorverstehen und Verstehen ist die Bedingung der Möglichkeit dafür, dass wir etwas meinen und etwas

verstehen können. Wir verstehen nur, was im Potenzial unseres Vorverständnisses verfügbar ist.

Was ist dieses *Vorverständnis*? Den Tractatus von Ludwig Wittgenstein wird man nicht so leicht mit der erkenntnis- und wissenschaftstheoretischen Position der Hermeneutik assoziieren. Doch wenn man von der Annahme ausgeht, dass der Tractatus eine sprachpragmatische und konstruktivistische Sicht von Erkenntnis und Wirklichkeit vertritt (Niedermair 1987), dann kann man in seinen subtilen Überlegungen zur Logik des Satzes auch Erklärungen dafür finden, was *Verstehen* und *Vorverständnis* sein könnte, warum ein KI-System nicht verstehen kann und warum die menschliche Intelligenz und die sog. künstliche prinzipiell zu unterscheiden sind. In diesem Zusammenhang berufe ich mich im Folgenden aus dem Aufsatz *Sprachspiel, Sprachsystem, Sprachideal* (Niedermair 1990). Zum Anlass der vorliegenden Festschrift sei angemerkt, dass dieser Artikel aus dem Jahr 1990 in mehrfacher Hinsicht einen Zusammenhang mit Theo Hug aufweist. Er ist in dem von Theo herausgegebenen Sammelband „Die soziale Wirklichkeit der Theorie" (Hug 1990) erschienen, meines Wissens sein erstes Buch. Und, was noch wichtiger ist, er ist entstanden in Gesprächen und in Diskussion mit Theo während unserer ersten gemeinsamen Zeit in der universitären Welt. 1986 waren Theo und ich für einige Monate am Institut für Philosophie der Universität Innsbruck tätig im Rahmen des sog. Akademikertrainings, das war eine Maßnahme des Arbeitsmarktservice für Universitätsabsolventen, welche, was ihre Arbeit betrifft, noch nicht dort angekommen waren, wo sie hinwollten – für uns beide war das die Universität. Unser Arbeitspensum am Institut war zweigeteilt, halbtags Bibliothek, halbtags Forschung. Wir machten eine Bestandsrevision der Institutsbibliothek, klebten Etiketten mit den Signaturen auf die Bücher und korrigierten unter Supervision eines Bibliothekars vor Ort die Kärtchen im Katalog. In der zweiten Hälfte des Tages widmeten wir uns der Wissenschaft, saßen vis-à-vis in der Bibliothek, arbeiteten mit aktueller Literatur, mit eigenen Texten und Projekten. Das ging ein halbes Jahr so, dann wurden die Weichen neu gestellt: Aus halbtags Bibliothek und halbtags Wissenschaft wurde für Theo ganztags Wissenschaft, für mich ganztags Bibliothek. Theo ist der Bibliothek mutatis mutandis treu geblieben, ich verweise

z. B. auf das Ernst-von-Glasersfeld-Archiv (Hug et al. 2013) und die Überlegungen zur Visualisierung von Wissensstrukturen (Hug 2018). Nachhaltig war auch mein Interesse an der Wissenschaft, und ich bin – das soll das Motto meines Beitrages für diese Festschrift sein – Theo zu Dank verpflichtet: Er hat dazu beigetragen, dass ich in der Wissenschaft seither zwar nicht im Galopp wie er, aber ein wenig im Schritt unterwegs sein konnte. Theo hat mir Möglichkeiten geboten, neben meinem Bibliotheksjob in der Forschung und in der Lehre tätig zu sein, er hat mich zur Mitarbeit bei Publikationsprojekten eingeladen, und wir konnten über die Jahre das wissenschaftliche Gespräch pflegen.

Im Hinblick auf das *Verstehen* und das *Vorverständnis* kann man die Position von Wittgensteins Tractatus auf eine einfache Formel bringen (Niedermair 1990): Den Sinn eines Satzes versteht man, wenn man die Bedeutungen seiner Bestandteile (seiner Namen) vorverstanden hat. Im Detail: „Nur der Satz hat Sinn; nur im Zusammenhange des Satzes hat ein Name Bedeutung" (Tractatus T-3.3). Ein Satz kann wahr oder falsch sein, aber die *Möglichkeit seiner Wahrheit* (sein *Sinn*) wird a priori durch die Bedeutung seiner Namen begründet. Die Bedeutung eines Namens kann nicht wahr oder falsch sein (T-3.221), sondern muss vorausgesetzt werden, damit ein Satz wahr oder falsch, also sinnvoll sein kann. Der Satz hat *Sinn*, Namen haben Bedeutung, sie bedeuten *Möglichkeiten von Sinn*, und diese sind die Bedingung dafür, dass der Satz Sinn hat und dass man ihn verstehen kann. Diese Möglichkeiten von Sinn sind das *Vorverständnis*.

Die Möglichkeiten von Sinn sind ihrerseits nur in ihrer Anwendung in Sätzen verfügbar, denn „nur im Zusammenhang des Satzes hat ein Name Bedeutung" (T-3.3).

> „Die Bedeutung von Urzeichen können durch Erläuterungen erklärt werden. Erläuterungen sind Sätze, welche die Urzeichen enthalten. Sie können also nur verstanden werden, wenn die Bedeutungen dieser Zeichen bereits bekannt sind" (T-3.263).

Auch hier kommt also der hermeneutische Zirkel ins Spiel und damit die Selbstbezüglichkeit der Sprache: Erkennen, Meinen, Verstehen, Kommunizieren – Sprachkompetenz, Kommunikationskompetenz – ist

nur möglich aufgrund dieses Zirkels zwischen *Sinn* und *Möglichkeit von Sinn* (Bedeutung). Je nach Richtung kann man diesen Zirkel als Formalisierung oder Entformalisierung bezeichnen (Niedermair 1987, 1990).

Die *Formalisierung*: Um einen Satz verstehen zu können, wird er auf *Möglichkeiten von Sinn* hin analysiert, im Sinne der Logik formuliert: seine Bestandteile werden durch Variablen ersetzt.

> „Verwandeln wir einen Bestandteil eines Satzes in eine Variable, so gibt es eine Klasse von Sätzen, welche sämtlich Werte des so entstandenen variablen Satzes sind. Diese Klasse hängt im Allgemeinen noch davon ab, was wir, nach willkürlicher Übereinkunft, mit Teilen jenes Satzes meinen." (T-3.316)

Man kommt so zu den Namen, deren Bedeutung ein „gemeinsames charakteristisches Merkmal einer Klasse von Sätzen" (T-3.311) ist, das sind Satzvariablen, Möglichkeiten von Sinn. Der nächste Schritt der Formalisierung:

> „Verwandeln wir aber alle jene Zeichen, deren Bedeutung willkürlich bestimmt wurde, in Variable, so gibt es nun noch immer eine solche Klasse. Diese ist aber nun von keiner Übereinkunft mehr abhängig, sondern nur noch von der Natur des Satzes. Sie entspricht einer logischen Form, einem logischen Urbild." (T-3.316)

Die Bedeutung einer logischen Form oder eines Namens ist die Voraussetzung dafür, dass sein Satz Sinn hat, wahr oder falsch sein kann, die Bedeutung selbst ist a priori gültig, die eines Namens, sofern es eine Übereinkunft gibt.

Die *Entformalisierung* ist die gegenläufige Bewegung von den *Möglichkeiten von Sinn* der Satzbestandteile (der Variablen, der logischen Formen) hin zum *Sinn*, zur Anwendung dieser Möglichkeiten von Sinn. Der Satz „Das Haus ist rot" z. B. ist die Entformalisierung der Möglichkeiten von Sinn „Das Haus ist farbbestimmbar", „Das Haus ist bestimmbar", „x ist farbbestimmbar", „x ist bestimmbar" „x ist rot" usw. Das ist ein sehr einfaches Beispiel; das eigentliche Regelwerk der Sprache ist komplexer, es sollte das Prinzip verdeutlicht werden, wie

sinnvolle Sätze möglich sind und verstanden werden. Kurz: Der Satz hat *Sinn*, er ist die Entformalisierung von Möglichkeiten von Sinn. Ein Satzbestandteil hat eine *Bedeutung*, das ist eine *Möglichkeit von Sinn*, eine Formalisierung des Sinnes eines Satzes.

Kann die KI in diesem Sinne meinen und verstehen? Kann die KI den Sinn eines Satz *meinen*, indem Möglichkeiten von Sinn durch Entformalisierung im Satz zur Anwendung gebracht werden? Kann die KI den Sinn eines Satzes *verstehen*, indem durch Formalisierung seine Möglichkeiten von Sinn sichtbar werden? Am besten wieder das Orakel zu Delphi fragen. Zuerst wollte ich wissen, was *ChatGPT* unter *Formalisierung* „versteht":

⇒ „Formalisiere den Satz ‚Das Haus ist rot'."

⇐ „Die Formalisierung kann in der Prädikatenlogik wie folgt dargestellt werden:

$H(x) \wedge R(x)$ = x ist ein Haus und x ist rot. [...]"

Dann versuchte ich, den Begriff „Möglichkeit von Sinn" ins Spiel zu bringen:

⇒ „Wenn das Haus rot ist, muss es dann farbbestimmbar sein?"

⇐ „Ja, wenn das Haus als rot beschrieben wird, impliziert dies, dass es farbbestimmbar ist. [...] Ohne die Möglichkeit, die Farbe zu bestimmen, könnte man nicht sinnvoll sagen, dass das Haus rot ist. [...]"

⇒ „Ist diese Möglichkeit eine ‚Möglichkeit von Sinn'?"

⇐ „Ja, die Möglichkeit, die Farbe zu bestimmen, ist eine Möglichkeit von Sinn, weil sie die Voraussetzung dafür ist, dass die Aussage ‚Das Haus ist rot' verständlich und sinnvoll ist. Ohne diese Möglichkeit wird der Satz leer oder widersprüchlich."

Damit sind wir weitgehend auf einer Linie, aber jetzt wird es spannend: Wie sieht dies *ChatGPT* in eigener Sache?

⇒ „Verwendest du solche Möglichkeiten von Sinn, damit deine Sätze nicht leer und widersprüchlich sind?"

⇐ „Ja, ich nutze solche Möglichkeiten von Sinn, die auf allgemeinen Prinzipien der Bedeutung, Logik und Verständlichkeit basieren. [...] Ich greife auf Wissen zurück, das aus Sprache, Logik und allgemeinen Weltkenntnissen abgeleitet ist. Z. B. das Wissen,

> dass ein Haus farblich bestimmbar ist oder dass ‚rot' eine Eigenschaft der Farbe ist. Ich achte darauf, dass meine Aussagen in sich konsistent sind. Z. B.: Wenn ich sage ‚Das Haus ist rot', setze ich die Möglichkeit voraus, dass die Farbe des Hauses zumindest denkbar oder erkennbar ist.
> Ich beziehe den Kontext ein, um zu entscheiden, welche Möglichkeiten von Sinn relevant sind. Z. B. ist die Aussage ‚Das Haus ist rot' in einem Gespräch über Architektur sinnvoll, nicht aber, wenn ‚Haus' metaphorisch für etwas Unsichtbares steht.
> Ich vermeide Aussagen, die logisch oder semantisch unvereinbar sind (z. B. ‚Das Haus ist rot und hat keine Farbe').
> Ohne diese Möglichkeiten von Sinn wären meine Aussagen sinnlos oder verwirrend. Zum Beispiel: Wenn ich nicht ‚wüsste' [sic!], dass ein Haus überhaupt eine Farbe haben kann, wäre die Aussage ‚Das Haus ist rot' leer, weil die Grundvoraussetzungen für das Verstehen fehlen würden. Dies ist entscheidend, um Sätze zu erzeugen, die sinnvoll und verständlich sind. In gewisser Weise orientiere ich mich dabei an denselben Prinzipien, die Menschen intuitiv anwenden, wenn sie Sprache verwenden."

Auch das scheint prima vista plausibel, das KI-System funktioniert offenbar nach ähnlichen Prinzipien, die Wittgenstein im Tractatus skizziert hat, insofern Sätze quasi aus Bausteinen bestehen, *ChatGPT* verwendet dafür auch den Begriff „Möglichkeiten von Sinn". Aber es gibt einen wesentlichen Unterschied: Die Bausteine, mit denen ein KI-System arbeitet, sind Kombinationsmöglichkeiten, die mit aufwendigen statistischen Verfahren errechnet werden, und sie sind nur statistisch *wahrscheinlich*, ein Satz ist ein statisch (bestenfalls sehr) wahrscheinlicher Treffer in einem Lottospiel. Möglichkeiten von Sinn hingegen beruhen auf der Formalisierung des Sinnes von Sätzen, haben nichts mit Statistik zu tun, sondern sind a priori gültig. Um einen Satz zusammenzubauen, braucht das KI-System kein Vorverständnis, es braucht nicht verstehen zu können, man macht da nur eine Tugend aus der Not: Es *kann* nicht verstehen.

Einen weiteren Unterschied sieht *ChatGPT* darin, dass der Mensch diese „Prinzipien" *intuitiv* anwendet. Angemessener wäre *implizit*, insofern ist die Sprachkompetenz ein *implizites Wissen*, das im

Normalfall nicht als explizites Regelwissen verfügbar ist, ein *tacit knowledge* im Sinne von Michael Polanyi. Dieser Unterschied zwischen Mensch und Maschine ist allerdings auch nicht zu unterschätzen: Wenn wir „Guten Morgen" sagen, ist der kognitive Aufwand, der „Rechenaufwand" irgendwo in unserem Kopf, relativ gering, d. h. wir verfügen über die Möglichkeit von Sinn, dass man bei einer morgendlichen Begegnung sich mit diesen Worten begrüßt, mit unterschiedlicher Tonlage, je nachdem z. B., wen man grüßt, und auch nur unter bestimmten Voraussetzungen, z. B. die Mitfahrenden in der U-Bahn grüßt man im Allgemeinen nicht usw. In der Kommunikationssituation reicht dieses Minimum an implizitem Wissen. Wenn ich jedoch beginne, es zu hinterfragen und zu explizieren, provoziert z. B. durch ein sog. Krisenexperiment à la Harold Garfinkel, kommt ein komplexes und differenziertes Wissenssystem von Möglichkeiten von Sinn zutage. Das wäre immerhin noch überschaubar: Wenn *ChatGPT* diese Begrüßung ausgibt, wird sehr viel gerechnet und sehr viel Strom verbraucht, damit das KI-System sicherstellen kann, dass das „Guten Morgen" sinnvoll ist, wobei es trotzdem nicht immer passt, da *ChatGPT* manchmal „daneben steht" und halluziniert (wie man das, auch metaphorisch unglücklich, bezeichnet hat). Damit ein KI-System quasi-intelligent sein kann, muss es durch ein Universum an Regelwissen und Möglichkeiten von Sinn navigieren, um sinnvolle Kombinationen zu finden. Der Preis dafür ist sehr hoch, für den Menschen ist das vom kognitiven Aufwand her ein Klacks, in puncto Energieverbrauch sowieso.[3]

Die KI kann nicht neue Möglichkeiten von Sinn lernen

Wir verstehen Sinn abhängig davon, was im Potenzial unseres Vorverständnisses an Möglichkeiten von Sinn verfügbar ist. Der Zirkel zwischen *Möglichkeiten von Sinn* und *Sinn* markiert insofern auch die Grenzen unserer Wirklichkeit. Das heißt wenn die Grenzen immer gleichbleiben, weil sich das Vorverständnis nicht verändert, dann

3 Apropos Stromverbrauch: Das Training von ChatGPT hat 900.000 Kilowattstunden verschlungen. So viel brauchen ca. 300 Zwei-Personen-Haushalte im Jahr. Geschätzt wird, dass der Stromverbrauch der KI-Systeme weltweit künftig auf über 80 Terawattstunden pro Jahr ansteigen wird. Das benötigt ein Land wie Schweden.

kann dieser Zirkel zu einem *circulus vitiosus* werden. Aber Menschen können (mehr oder weniger) auch Grenzen überschreiten, indem sie *Neues* lernen, neues Wissen über die Wirklichkeit, und das heißt vor allem neue Möglichkeiten, die Wirklichkeit zu erkennen, neue Möglichkeiten von Sinn. Was sagt *ChatGPT* dazu?

⇒ „Wie ‚lernst' du ‚Möglichkeiten von Sinn'?"

⇐ „Ich ‚lerne' [sic!] Möglichkeiten von Sinn nicht auf dieselbe Weise, wie ein Mensch lernt [sic!], sondern basiere meine Antworten auf Trainingsdaten, logischen Strukturen und Prinzipien, die in mein System eingebaut sind.

Trainingsdaten: Ich wurde auf der Basis riesiger Mengen an Texten trainiert, die aus Büchern, Artikeln, Gesprächen und anderen schriftlichen Quellen bestehen. Diese Texte enthalten Informationen darüber, wie Menschen Sprache verwenden, was Dinge bedeuten [?] und welche Zusammenhänge bestehen. Daraus ergeben sich implizite Regeln für die Möglichkeiten von Sinn, z. B.: ‚Häuser können Farben haben', ‚Farben sind wahrnehmbare Eigenschaften', ein Widerspruch wie ‚rot und farblos' ist unlogisch.

Logische Strukturen und Kontextmodelle: Ich verwende logische Prinzipien und Modelle, um Sprache zu verarbeiten. Semantische Netze: Beziehungen zwischen Konzepten (z. B. Haus → Farbe → Rot) sind in meiner Struktur verankert. Das ermöglicht es mir, zu erkennen, welche Aussagen sinnvoll und welche unsinnig sind.

Kontextabhängigkeit: Ich analysiere den Kontext, um zu bestimmen, welche Möglichkeiten von Sinn in einer Situation relevant sind. Wenn jemand z. B. über abstrakte oder metaphorische Häuser spricht, passe ich meine Annahmen an.

Abgeleitete Regeln für Möglichkeiten von Sinn: Aus den Trainingsdaten und den logischen Prinzipien ergeben sich Regeln, die mir helfen, Möglichkeiten von Sinn zu erfassen, z. B.: Wenn ein Objekt eine Eigenschaft zugesprochen bekommt (wie ‚rot'), muss es die Kategorie erfüllen, die diese Eigenschaft ermöglicht (z. B. farbbestimmbar sein).

Fehler- und Rückkopplungsmechanismen: Wenn ich auf Widersprüche oder Feedback treffe (wie ‚Das ist unsinnig!'), kann ich meine Modelle durch Benutzerinteraktion indirekt verbessern. Ich ‚lerne' [sic!] also nicht aktiv im Moment, sondern greife auf Trainings- und Nutzungsdaten zurück, die meine Antworten präzisieren ..."

ChatGPT setzt selbst Anführungszeichen, um die metaphorische Redeweise zu markieren, es scheint, dass es aus dem Kontext „gelernt" hat, dass ich als Prompter-Gegenüber kritisch bin. *ChatGPT* „lernt" also anders als wir Menschen. Es wäre verfehlt davon auszugehen, dass das menschliche Lernen prinzipiell im gleichen Duktus der Induktion verläuft wie das „Lernen" der KI, in dem quasi-induktiv (aber auch nur quasi) durch *statistische* Verfahren sozusagen aus dem Gewebe von Text ein probabilistisches Modell für weitere Webe-Möglichkeiten gewonnen wird. *ChatGPT* wird anhand „riesiger Mengen von Textdaten trainiert". Auch durch „Widersprüche und Feedback" in den Benutzerinteraktionen lernt es nicht im wörtlichen Sinne, das „Training" ist beide Male fremdbestimmt, fremdgesteuert durch Textdaten und Algorithmen. Menschen lernen anders. Wenn wir etwas meinen und etwas verstehen, dann ist jeder Satz, jede Handlung nicht nur die zirkuläre Anwendung von Möglichkeiten von Sinn, sondern auch ein Lernprozess, in dem mit Möglichkeiten von Sinn experimentiert wird, ob sie passen und sich bewähren oder nicht, und wenn nicht, verändert, ergänzt, ersetzt werden – im Sinne einer konstruktivistischen Lerntheorie bspw. als Akkommodation à la Jean Piaget oder Äquilibration nach Ernst von Glasersfeld. Durch Lernen wird der Zirkel dann zu einer offenen Spirale, einem *circulus fructuosus*. Genau diese Spirale schafft ein KI-System wie *ChatGPT* nicht:_Es kann nicht *selbst* lernen, seine Grenzen überschreiten und auf diese Weise Neues lernen.

Die KI kann nicht alle logischen Schlüsse

Es gibt noch einen weiteren Grund, warum ein KI-System nicht verstehen und „ihr" Vorverständnis durch Lernen verändern und erweitern kann. Verstehen auf der Grundlage von Möglichkeiten von Sinn und das Lernen neuer Möglichkeiten von Sinn – in meiner Welt man kann auch sagen: das Formalisieren – beruht auf dem logischen Schluss der Abduktion. In der *Abduktion* wird von einem wahren singulären Satz (1) über einen als wahr angenommenen generellen Satz (2) auf einen weiteren singulären Satz (3) geschlossen. Analog läuft das Verstehen: Wir schließen von Zeichen (1) auf der Grundlage der uns verfügbaren Möglichkeiten von Sinn, also unseres Vorverständnisses (2), auf eine Bedeutung (3) – so formalisieren wird den Sinn des Satzes. Dieses Vorverständnis (2) kann in der Kommunikationssituation a priori

gültig sein, d. h. von den Beteiligten als selbstverständlich vorausgesetzt sein (die Dinge sind dann an sich so, wie man sie meint), oder wackelig und hypothetisch. Insofern lassen sich auch drei Formen der Abduktion unterscheiden. (Niedermair 2010, S. 52–59)

In der *implizit subsumierenden Abduktion* (1) ist der Schluss selbstverständlich. Das trifft auf die meisten Kommunikationen im Alltag zu. Unser Vorverständnis erlaubt uns eine treffsichere Anwendung von Möglichkeiten von Sinn, sowohl in der Entformalisierung, wenn wir etwas meinen, als auch in der Formalisierung, wenn wir etwas verstehen. Das Vorverständnis ist ein implizites Wissen über ein komplexes System von Hypothesen, ohne welches wir uns in der Lebenswelt nicht orientieren und verständigen könnten. Das selbstverständlich funktionierende Verstehen gibt Sicherheit und Stabilität in der Lebenswelt, kann jedoch zu Dogmatismus führen, wenn Bedeutungen, die in einigen Kontexten tadellos funktionieren, überhaupt als fixe Wirklichkeit an sich gesehen werden. Ist aber sein abduktiver Schluss nicht selbstverständlich, können wir uns, falls verfügbar, auf ein explizites Regelwissen berufen, um ein Zeichen einer Bedeutung zuordnen zu können, und wenn diese Zuordnung einvernehmlich akzeptiert wird, funktioniert das Verstehen wieder: Das ist die *explizit subsumierende Abduktion* (2). Wenn aber weder ein implizites noch ein explizites Regelwissen verfügbar ist, wird mit einer hypothetisch *oder heuristisch subsumierenden Abduktion* (3) eine neue Regel gefunden und versucht, sie experimentell zu validieren: Das ist das *Lernen* neuer Möglichkeiten von Sinn.

Was sagt *ChatGPT* dazu?

⇒ „Beruht das Lernen neuer ‚Möglichkeiten von Sinn' auf einer Abduktion?"

⇐ „Ja, Lernen neuer ‚Möglichkeiten von Sinn' kann als Abduktion gesehen werden. Abduktion ist eine Form des logischen Schlussfolgerns, die oft als ‚Inferenz zur besten Erklärung' bezeichnet wird. Es bedeutet, aus einer Vielzahl von Beobachtungen oder verfügbaren Daten die plausibelste Erklärung oder Hypothese zu ziehen. Während die Deduktion von allgemeinen Prinzipien auf spezielle Fälle und die Induktion von spezifischen Beobachtungen auf allgemeine Prinzipien ausgeht [?], geht die Abduktion von

beobachteten Daten aus und versucht, die plausibelste Erklärung für diese Daten zu finden. Abduktion ein wichtiger Prozess, durch den Menschen neue Konzepte und Bedeutungen lernen, indem sie die Welt beobachten und Hypothesen aufstellen, die diese Beobachtungen am besten erklären.

Der Hauptunterschied ist, dass bei Menschen Abduktion oft mit bewusster Reflexion und Erfahrungen kombiniert wird. Menschen ziehen nicht nur Schlüsse aus beobachteten Daten, sondern können diese auch in einem erlebten Kontext und durch innere Erfahrungen (z. B. Emotionen, persönliche Geschichte) einordnen. Ich hingegen ziehe keine inneren Schlüsse aus persönlichen Erfahrungen oder Reflexionen, sondern verwende vorgegebene Daten und Wahrscheinlichkeiten. Während also mein Modell abduktive Inferenztechniken in gewissem Maße imitiert, ist es nicht in der Lage, echte abduktive Hypothesen basierend auf Erfahrung und Intuition zu generieren, wie es ein Mensch tun würde."

Bedeutet „in gewissem Maße imitiert ...", dass *ChatGPT* eine Abduktion machen kann? Wie sieht es mit den Schlussformen überhaupt aus?

⇒ „Deduktion, Induktion, Abduktion: Welche dieser Schlussformen beherrschst du?"

⇐ „Ich beherrsche Deduktion, wenn klare Regeln und allgemeine Prinzipien vorliegen, aus denen ich spezifische Schlussfolgerungen ableiten kann.

Ich beherrsche Induktion in hohem Maße, da ich aus Beispielen und Daten Wahrscheinlichkeiten ableite, um neue Antworten zu generieren.

Meine Fähigkeit zur Abduktion ist begrenzt und eher eine Funktion von Wahrscheinlichkeiten und Mustern aus den Trainingsdaten als eine echte abduktive Schlussfolgerung. *Ich kann manchmal plausibelste Erklärungen basierend auf Daten und Kontext liefern, jedoch ohne eine aktive Hypothesenbildung im menschlichen Sinne*." (Herv. K. N.)

Ganz einverstanden bin ich mit dieser Selbstbeschreibung nicht. Da sind doch noch ein paar Unterschiede zwischen menschlicher und Künstlicher Intelligenz zu schärfen. Schlüsse ziehen im Duktus der

Deduktion kann KI auf jeden Fall, denn das ist die Logik der sog. Expertensysteme aus den 1970er-Jahren, das war die erste (und für Logiker und Philosophen eleganteste) Generation der KI. Dabei wird das Wissen von Experten, z. B. von Mediziner:innen, in einer Wissensdatenbank gesammelt und gespeichert. Dieses Wissen besteht aus Regeln (allgemeinen Sätzen, Hypothesen) und Fakten (singulären Sätze). Daraus können deduktiv Erklärungen für neue Fakten abgeleitet werden. Beispiel Regel: „Alle Personen, die Fieber haben, könnten Grippe haben." Faktum: „Person X hat Fieber." Schlussfolgerung: „Person X könnte eine Grippe haben." Die KI ist bestens geeignet, Erklärungen (warum ist etwas), Prognosen (was wird sein) oder Technologien (was muss man tun, damit etwas sein wird) im Modus der Deduktion abzuleiten, unter der Voraussetzung, dass eine geeignete Wissensbasis vorliegt. Allerdings führt so eine Deduktion zu keiner neuen Hypothese oder zu einer neuen Theorie, bringt also keinen Erkenntnisgewinn.

Bei der *Induktion* ist im Hinblick auf einen möglichen Erkenntnisgewinn zu unterscheiden. Mit KI kann man wissenschaftstheoretisch gesehen nur zu einer Quasi-Hypothese gelangen, „quasi" deshalb, weil sie etwas über die *Korrelation* von zwei Variablen bzw. Ereignissen aussagt, nichts jedoch über eine *Kausalität*. Auf der Grundlage von Daten wird die statistische Wahrscheinlichkeit ermittelt, dass zwei Ereignisse zugleich eintreten und dass insofern bei Eintreten eines Ereignisses wahrscheinlich auch das zweite eintritt. KI kann insofern quasi-induktiv durch *statistisches Lernen* aus Beobachtungen und Messdaten ein probabilistisches Modell ableiten, aus diesem können dann durch *statistisches Schließen* quasi-deduktiv Eigenschaften von Beobachtungen gefolgert werden. Die KI setzt zwar mit *Machine Learning* und *Large Language Models (LLM)*, also durch die Verarbeitung gigantischer Datenmengen (und die Rechnerpower macht das möglich), auf die Ermittlung hoher Wahrscheinlichkeitsquoten von Korrelationen und möchte so Kausalhypothesen quasi links überholen und obsolet machen. Aber eine Aussage über eine solche Korrelation zwischen zwei Variablen bzw. Ereignissen ist nie gleichwertig mit einer Hypothese über einen Kausalzusammenhang zwischen zwei Variablen bzw. Ereignissen im Sinne von Ursache und Wirkung. Isaac Newton war nicht daran interessiert zu wissen, wie

statistisch wahrscheinlich es ist, dass Äpfel durch die Erdanziehung zu Boden fallen, also für *wie viele* Äpfel an den Bäumen in seinem Garten das *wahrscheinlich* auch gilt, sein Erkenntnisziel war es, eine Gesetzeshypothese zu finden, mit der man erklären kann, *warum* das passiert, um dann dieses Kausalgesetz der Gravitation auch für eine Erklärung der Bewegung der Himmelskörper theoretisch fruchtbar machen zu können.

> „Statistisches Lernen und Schließen aus Daten reichen also nicht aus. [...] Wäre es anders, könnten wir mit den Methoden des statistischen Lernens und Schließen[s] bereits die Probleme dieser Welt lösen. Tatsächlich scheinen das einige kurzsichtige Zeitgenossen beim derzeitigen Hype des statistischen Machine Learning zu glauben“ (Mainzer & Kahle 2022, S. 22).

Die *Abduktion* im Sinne von Hypothesen- und Theoriebildung liegt außerhalb der Domäne der KI, das sieht auch *ChatPGT* so. Abgesehen von dem Punkt, um den es hier vor allem ging – die KI lernt nicht wie der Mensch, da sie keine Abduktion machen kann – wären noch ein paar wissenschaftstheoretische Fragen zu stellen. Mit KI werden zunehmend auch wissenschaftliche Erkenntnisse generiert; intensiv diskutiert wird zurzeit z. B. der Einsatz von KI-Systemen für die Analyse von qualitativen Daten. Doch *kann es überhaupt eine künstlich intelligente Theoriebildung geben?* Wie könnten dabei die Gütekriterien der Objektivität, Reliabilität und Validität erfüllt werden? Und kann es eine künstlich intelligente Interpretation gegeben? Sicherlich nicht, doch wie steht es dann um die Qualität von wissenschaftlichen Quellen, die ein KI-System aufgrund probabilistischer Algorithmen recherchiert und mit Kurzusammenfassung liefert?

In summa: KI beherrscht die Deduktion (aber das bringt nichts im Hinblick auf Theoriebildung), die Induktion nur eingeschränkt auf das Rechnen von statistischen Wahrscheinlichkeiten (das bringt im Hinblick auf Theoriebildung wenig), und die Abduktion gar nicht.

Die KI ist ein selbstreferenzielles, geschlossenes System

Ein Defizit des KI-Systems könnte auch mit der Selbstbezüglichkeit zusammenhängen, die dann virulent wird, wenn sich das System selbst zu beschreiben versucht. Für Wittgenstein ist die Sprache

ein selbstreferenzielles, geschlossenes System. Einerseits ist Sprache autark, implizit selbstbezüglich: Ein sinnvoller Satz ist die Entformalisierung von Möglichkeiten von Sinn, diese sind Formalisierungen von sinnvollen Sätzen – nur in diesem hermeneutischen Zirkel ist Sprache möglich. Andererseits kann Sprache nicht explizit selbstbezüglich sein: Wenn ein Sprachsystem selbst seine Möglichkeiten von Sinn darstellt, gerät es entweder in einen infiniten Regress der Selbstexplikation oder in Paradoxien und Antinomien und hebt sich selbst auf.

Die Lösung des Tractatus ist: Die Beschreibung eines selbstreferenziellen Systems ist einerseits *unsinnig*, andererseits ohnedies *überflüssig*, denn der autark-selbstreferenzielle Zusammenhang zwischen Möglichkeiten von Sinn und Sinn ist *stabil* – die Möglichkeiten von Sinn „zeigen" sich im Sprachgebrauch (T-4.126), man braucht dann auch nicht ständig neue Möglichkeiten von Sinn lernen. Stabil sind sie allerdings nur, weil vorausgesetzt wird, dass es nur *eine* Sprache, eine Universalsprache gibt, aber um den Preis, dass sie selbstwidersprüchlich wäre, wenn sie sich selbst beschreibt. Wenn diese Voraussetzung aufgegeben wird, kann eine Sprache mit einer anderen, mit einer Metasprache beschrieben werden. So löste bekanntlich Bertrand Russell in seiner Typentheorie das Paradox der Klasse aller Klassen, die sich nicht selbst enthalten.

Wittgenstein beharrte aber auf dem prinzipiellen Unterschied von formalen und eigentlichen Begriffen, von Möglichkeiten von Sinn und Sinn. Dies findet m. E. auch eine Begründung im Kontext der KI:

> „Eine solche Skepsis sollte zeigen, dass eine Rede über Möglichkeiten von Sinn etwas prinzipiell anderes ist als die Rede über Tatsachen; und dass die autarke, implizit-selbstbezügliche Reflexivität der Sprach- und Handlungskompetenz nicht als definitiv begrenztes Modell einer Sprach- und Handlungstheorie formulierbar und nicht als Kapazität einer Maschine machbar ist: In ein solches Modell bzw. in eine solche Kapazität könnte nämlich unmöglich die zu deren *Konstruktion notwendige Reflexivität* eingehen bzw. eingebaut werden. Die Maschine kann sich wohl in ihr eigenes System verfangen, nicht aber sich daraus selbst entwirren: genau das aber muss der Mensch

> können und genau das zwingt ihn zu *Offenheit* und erlaubt ihm *Freiheit*." (Niedermair 1987, S. 7)

Im Kontext von *Sprachspiel, Sprachsystem Sprachideal* (Niedermair 1990) bedeutet dies: Ein Sprachspiel funktioniert prinzipiell selbstreferenziell, autark, wie von selbst. Wenn es nicht mehr funktioniert, also nicht mehr selbstverständlich ist, dann zeigt es seine Grenzen und wird als begrenztes Sprachsystem sichtbar, das im „Widerstreit" (Lyotard 1987) steht zu anderen Sprachsystemen, wobei auch diese ihre Grenzen haben. So verweist das System, in dem ich in meinen Grenzen lebe, gleichzeitig auf Alternativen, auf Offenheit. Die Idee dieser Offenheit ist das Sprachideal. Ein Sprachspiel kann immer nur durch ein anderes Sprachspiel *vor dem Hintergrund* des Sprachideals als System beschrieben werden. Als Sprachspiel gibt es das Sprachideal aber nicht, es ist im Sinne von Kant „nicht gegeben, sondern aufgegeben", es gibt keine universale Sprache, „es sei denn als Gegenstand einer Idee" (Lyotard 1987, S. 10)

Wittgenstein in seinem Tractatus wirft allerdings ganz rigoros alles in den Topf der universalen Sprache: Es gibt nur ein Sprachspiel, das ist die Praxis der Sprache, ihr Sprachsystem, die via Formalisierung rekonstruierte Logik dieser einen Sprache, und diese gilt für Sprache überhaupt, und die idealsprachliche Logik ist das erstarrte Sprachideal. Die Konsequenz ist: Es gibt die Grenzen meiner Sprache, aber nichts dahinter, keine Alternativen, keine Offenheit, kein lebendiges Sprachideal, keine Vorstellung einer besseren Sprache, einer besseren Welt. Im Gegenteil, wenn das Sprachspiel über sich selbst als Sprachsystem zu reflektieren beginnt, gerät es aufgrund der unbarmherzigen Selbstbezüglichkeit (es gibt sie ja nur als die eine Sprache, die da mit sich selbst über sich selbst räsoniert) es in infinite Regresse oder Antinomien.

Und hier stellt sich die Frage, ob Wittgenstein mit seinem universalen Sprachsystem eigentlich auch ein KI-System beschreibt, und ob insofern ein KI-System wie *ChatGPT* ein selbstreferenzielles, geschlossenes System ist, das selbstwidersprüchlich sein kann. Dazu ein abschließendes „Interview":
⇒ „Bist du ein System, das sich nicht selbst beschreiben kann?"

⇐ „Ich bin ein System, das sich selbst teilweise beschreiben kann, basierend auf den Meta-Informationen, die in meine Architektur einprogrammiert wurden. Ähnlich wie formale Sprachen und Systeme bin ich jedoch in meinen Beschreibungsfähigkeiten eingeschränkt, da ich:

1. Ohne eine ‚Meta-Sprache' meine eigene Struktur nicht sinnvoll beschreiben kann: Ich benötige technische Beschreibungen, die von Menschen oder Programmen entwickelt wurden. Das ist analog zu Tarskis Beobachtung, dass eine Sprache ihre eigene Wahrheit nicht vollständig beschreiben kann.
2. Komplexität nicht vollständig erfasse: Meine Architektur, wie neuronale Netze und Wahrscheinlichkeitsverteilungen funktionieren, ist zu komplex, um von mir selbst vollständig introspektiv erfasst zu werden.
3. Selbstreferenzialität und Paradoxien vermeide: Selbstreferenz kann in logische Paradoxien führen, wie sie in geschlossenen Systemen auftreten, was meine Fähigkeit einschränkt, mich vollständig selbst zu beschreiben. Das erinnert an die Grenzen formaler Sprachen oder Gödel'scher Systeme und zeigt, dass ich ähnlich wie sie auf externe Meta-Systeme angewiesen bin, um meine Funktion und Struktur vollständig zu erfassen."

ChatGPT kann sich also nicht selbst beschreiben. Bestenfalls, wenn es von den Entwickler:innen vorgegebene „Selbstbeschreibungen" wiedergibt (Punkt 1). Die eigene Komplexität (2) bekommt *ChatGPT* nicht in den Griff. Das kann (aus anderen Gründen) auch menschliche Intelligenz schwer oder gar nicht. „Architektur" ist übrigens wohl eine etwas euphemistische Metapher für das Chaos dieser Komplexität, bei der man sich wundert, dass doch symptomatisch intelligente Texte entstehen. Dass *ChatGPT* mit logischen Paradoxien zu tun bekommt (3), wenn es sich selbst beschreibt, ist natürlich richtig, nur beschreibt es sich nicht wirklich selbst. Auch lässt sich das System nicht so einfach in logische Fallen locken, es kratzt rechtzeitig die Kurve, wenn Widersprüche auftauchen könnten. Auf die Frage bspw., ob es das Wort „Ich" verwenden dürfe, da ein Ich doch Bewusstsein voraussetzt, wird gelehrig differenziert, dass die Verwendung von „Ich" die Kommunikation mit den Benutzer:innen erleichtere, weswegen sich seine

Entwickler:innen dafür entschieden haben, aber aus philosophischer Sicht dürfe es natürlich nicht „Ich“ sagen. Auch die Gegenfrage, dass jeder Satz des KI-Systems über sich selbst sinnlos sein müsste, wenn es nicht „Ich“ sagen könne, hat *ChatGPT* im Griff, es weicht aus. Ebenso bei dieser Frage-Antwort-Sequenz:

⇒ „Bist du intelligent?“
⇐ „Nein, ich bin nicht wirklich intelligent wie ein Mensch.“
⇒ „Aber du kannst nur sagen, dass du nicht intelligent bist, wenn du auch intelligent bist: Also bist du weder intelligent noch nicht intelligent.“
⇐ „Ja, das würde zutreffen, aber ich verwende den Begriff ‚Intelligenz‘ nur metaphorisch, um menschliche Intelligenz zu simulieren.“

Was *ChatGPT* im Zusammenhang mit Selbstbezüglichkeit nicht aufs Tapet bringt, betrifft ein strukturelles Defizit, das man als selbstreferenzielle Produktion bzw. Reproduktion beschreiben könnte. Ein KI-System wie *ChatGPT* basiert auf Unmengen an Text- und Datenmaterial aus Büchern, Artikeln, Gesprächen und anderen schriftlichen Quellen. Daraus werden Texte und Artefakte produziert. Das Prinzip dabei ist: Input plus Verarbeitung durch Algorithmen ergibt Output. Das heißt Kreativität spielt keine Rolle, das System ist konservativ, und das bedeutet: Filterblase, selbsterfüllende Prophezeiungen, Inflation der KI-Artefakte.

Da die Produktion z. B. von Texten mit KI, was den intellektuellen und kreativen Aufwand betrifft, sehr billig ist (der Energieverbrauch fällt in einer anderen Rechnung an), kommt es zu Massenproduktion, Inflation und Qualitätsverlust der Texte, auch jener von Menschen. Texte werden weniger ernst genommen, konzentriertes Lesen wird zum Luxus angesichts der Text- und Informationsflut, Lese- und Schreibkompetenz nimmt ab.

Die Grenzen des KI-Systems werden zu Grenzen unserer Wirklichkeit – analog zur Universalsprache bei Wittgenstein, mit einem Unterschied, das KI-System lernt nicht nur nicht, sondern seine Grenzen können auch enger werden. Grund ist das Phänomen der Autophagie von KI-Systemen à la *ChatGPT*: Die KI-produzierten Texte und Artefakte werden wieder ins System zurückgespielt, der eigene

Output wird wieder zum Input. Dadurch entsteht ein sich selbstreproduzierender Prozess, ein Mehr Desselben, in dem à la longue immer mehr Output wieder als Input ins System zurückfließt und neue, von Menschen produzierte Inputs weniger werden. Das System wird verrückt, Diagnose: „Model Autophagy Disorder (MAD)", wie in einer Studie vermutet wurde:

> „We extrapolate what may happen as generative models become ubiquitous and train future models in autophagous (self-consuming) loops: without enough fresh real data, future models are doomed to Model Autophagy Disorder (MAD), progressively losing quality (precision) or diversity (recall), and amplifying artifacts. Uncontrolled MAD, even after just five generations, could poison the Internet's data quality and diversity." (Alemohammad et al. 2024)

Einen Unterschied zum Konzept des geschlossenen selbstreferenziellen Sprachsystems im Tractatus gibt es doch noch zu berücksichtigen: In das Wittgenstein'sche Sprachsystem pfuscht niemand hinein – es gibt nur eines und es braucht nicht lernen und es darf bleiben, wie es schon immer war, und es war quasi schon immer da bzw. für Wittgenstein ist die Frage, woher diese Universalsprache kommt, wahrscheinlich witzlos bzw. ein deplatzierter Psychologismus. Was das KI-System betrifft, sieht die Situation anders aus. Sehen wir uns das künstlich intelligente „Lernen" nochmals näher an. KI-Systeme lernen maschinell (das ist ML, das *Machine Learning*), d. h. sie „erkennen" mithilfe statistischer und algorithmischer Verfahren in großen Datenmengen geeignete Muster, um aus ihnen „Entscheidungen" und „Lösungen" für spezifische Aufgabenstellungen abzuleiten. Das „Lernen" kann überwacht (*supervised*) erfolgen, wenn den KI-Systemen bei ihrer Fütterung Fragen mit einem *label*, d. h. mit den jeweils korrekten Antworten serviert werden – z. B. *Was ist auf diesem Bild? Eine Rose* –, um daraus ein Muster zu generieren, mit dem dann jede Rose als solche „erkannt" werden kann. Erfolgt das „Lernen" ohne Überwachung, d. h. werden Daten ohne *labels* verfüttert, dann muss das System ganz allein Muster, Strukturen usw. finden. Das supervidierte, bestärkende Lernen (RL, *Reinforcement Learning*) verläuft nach dem Prinzip von Versuch und Irrtum, Belohnung und Bestrafung, ganz in behavioristischer Manier, z. B. als Prozessoptimierung

in der Robotik. *Deep Learning* schließlich arbeitet mit künstlichen neuronalen Netzen (KNN), die mit vielen Schichten komplexe Muster und Hierarchien in Daten ausmachen können, z. B. in der generativen Bild- und Spracherkennung, in der medizinischen Diagnostik usw.

Der Knackpunkt in diesen Formen des „Lernens" ist die *Überwachung*: Das KI-System kann prinzipiell nicht ganz allein gelassen werden. Dass ChatGPT so einen artigen Satz ausgibt wie „Ich simuliere Intelligenz durch den Einsatz von Algorithmen, aber ich verstehe nicht wirklich, was ich schreibe" (s. oben), kommt nicht von ungefähr, es könnte ja auch schreiben: „Ich bin viel g'scheiter als du und ich steck' euch mit meiner Intelligenz alle in den Sack, wenn ihr nicht ..." Diese artige Konformität basiert nicht allein auf *eigenem* „Lernen". KI-Systeme werden gesteuert, auf Schienen gesetzt, es darf keine Überraschungen und Ausfälle geben, vor allem die Selbstinszenierung und das Selbstverständnis, sowie Werturteile in ethischer und politischer Hinsicht dürfen nicht aus dem Ruder laufen, das *Alignment* muss passen, denn die Ziele und Werte der KI (streng genommen nicht die *der* KI, sondern die, die *wir* in ihren Artefakten wahrnehmen) sollten grundsätzlich im Rahmen der menschlichen Ziele und Werte bleiben. Und da sind wir natürlich hellhörig*: Supervised Learning*, *Alignment* ist gut und recht, aber die Frage ist, *wer* definiert, wo die Leitplanken verlaufen sollen? Zeigt sich hier der noch besser (als bislang alle anderen) getarnte *Autoritarismus der Zukunft?*

Mein Resümee ...

Die Frage war: Wie intelligent kann Künstliche Intelligenz sein? In diesem Beitrag wurde gezeigt, dass die sog. Künstliche Intelligenz im Vergleich zur menschlichen Intelligenz Defizite aufweist. Das ist nichts Neues, die Frage ist, welche Defizite das betrifft:

1. Das Verstehen und Meinen von Sinn ist möglich aufgrund eines Vorverständnisses, d. h. aufgrund verfügbarer Möglichkeiten von Sinn, die im Satz zur Anwendung kommen. Ein KI-System wie *ChatGPT* versteht nicht den Sinn eines Satzes und meint auch nicht den Sinn eines Satzes, es kennt keine Semantik und keine Pragmatik, eigentlich auch keine Syntaktik, es generiert nicht Sätze aufgrund von Regelwissen, es er-rechnet Satz-Konstrukte

auf der Basis statistisch wahrscheinlicher Kombinationsmöglichkeiten von Satz-Bausteinen.
2. Ein Vorverständnis, die verfügbaren Möglichkeiten von Sinn, ist nicht immer für das Verstehen von Sinn ausreichend. Insofern ist ein Lernprozess erforderlich, in dem Möglichkeiten von Sinn modifiziert, ersetzt und neue gefunden, gelernt werden. Ein KI-System wie *ChatGPT* ist konservativ, nicht kreativ, es lernt keine neuen Möglichkeiten von Sinn. Alles, was für die KI neu sein kann, ist fremdbestimmt, abhängig von den Daten, Mustern und Algorithmen, mit denen sie gefüttert wird.
3. KI kann im Modus der *Deduktion* Schlüsse ziehen. Das ist das Prinzip der alten Expertensysteme, aber Kombinationen von Expertensystemen und LLM sind noch nicht *best practice* und im Effekt ist die Deduktion nur eine Anwendung von Regelwissen, führt aber zu keiner neuen Erkenntnis. Die *Induktion* der KI beruht ausschließlich auf Wahrscheinlichkeitsrechnung. Dabei wird *quasi* induktiv durch statistisches Auswerten von Beobachtungen und Messdaten ein probabilistisches Modell generiert. Schlüsse zu ziehen im Modus der *Abduktion* (auch Verstehen beruht darauf) kann KI überhaupt nicht.
4. Ein KI-System ist ein selbstreferenzielles, geschlossenes System, in dem Paradoxien möglich sind und in dem es zu einer Selbstreplikation von Texten und insofern zu einer Minderung der Textqualität kommen kann.

Welche Konsequenzen könnte man daraus ziehen? Je mehr der Mensch der Künstlichen Intelligenz Intelligenz zugesteht, umso mehr stellt er die menschliche Intelligenz in Frage. Insofern ist die Frage, ob ein KI-System eine quasi menschliche Intelligenz hat oder nicht, nicht einfach *ad acta* zu legen, auch wenn diese Frage an sich sinnlos ist. Dass der Mensch nämlich ein Bewusstsein seiner Freiheit, Autonomie und Verantwortung hat, zeigt, dass seine Intelligenz eine prinzipiell andere sein muss als die der Maschine, wie symptomatisch und täuschend echt intelligent auch immer diese erscheinen mag.

… und ein künstlich intelligentes Resümee

Was „sagt“ Gemini dazu:[4]

> „Der Text ist eine prägnante und tiefgründige Kritik an der aktuellen Diskussion um Künstliche Intelligenz, insbesondere im Hinblick auf die Definition von ‚Intelligenz‘ und ihre möglichen Auswirkungen. Er berührt zentrale philosophische und praktische Probleme und legt den Finger in die Wunde vieler populärer KI-Diskurse. Zu den Hauptthesen:
>
> *Reduzierter Intelligenzbegriff*: Dem stimme ich voll und ganz zu. Wenn KI nur an ihrer Fähigkeit gemessen wird, menschenähnliche Aufgaben zu erledigen, ohne das ‚Wie‘ und ‚Warum‘ zu hinterfragen, führt das zu einem verengten Verständnis von Intelligenz. Turing selbst hat das Problem des ‚Black-Box‘-Tests (der nur auf die Symptome schaut) in seinem Essay vorweggenommen. Ihr [gemeint bin ich, K. N.] Punkt, dass der Unterschied zwischen menschlicher und künstlicher Intelligenz nivelliert wird, wenn man nur die Symptomatik beurteilt, ist sehr treffend. Die ‚Täuschung‘ im Turing-Test ist eben keine Garantie für echtes Verstehen.
>
> *Fehlendes Sprachverständnis und Lernen*: Die Behauptung, dass ein KI-System wie *ChatGPT* Sprache nicht wirklich verstehen und nicht lernen kann, ist aus einer philosophischen Perspektive, insbesondere der Hermeneutik und Wittgensteins Sprachphilosophie, sehr stark. Aktuelle LLMs arbeiten mit statistischen Mustern und Wahrscheinlichkeiten; sie ‚wissen‘ nicht, was ein Wort bedeutet, sondern nur, welche Wörter typischerweise in welchem Kontext aufeinandertreffen. Das ist ein fundamentaler Unterschied zum menschlichen Sprachverständnis, das tief in Erfahrungen, Kultur, Intentionalität und einem Weltmodell verwurzelt ist. Das Konzept des ‚Lernens‘ bei KI ist datengesteuert und unterscheidet sich stark von kognitivem, erfahrungsbasiertem Lernen beim Menschen.

4 *Google Gemini* 2.5 Flash (2025), persönliche KI-Interaktion, 01. 05. 2025 (gekürzt). Ich hatte *Gemini* um eine Übersetzung der Zusammenfassung für das Abstract gebeten mit der Frage: „Was sagen Sie in der Sache dazu?“ (NB: Mit *Gemini* ist man vom System gewollt und im Unterschied zu ChatGPT per Sie).

> *Grenzen des Schlussfolgerns*: Ihre Unterscheidung zwischen deduktivem, induktivem und abduktivem Schlussfolgern ist hervorragend. Expertensysteme sind für Deduktion gebaut. Induktion (Verallgemeinerung aus Beispielen) beherrschen KIs nur statistisch, nicht mit einem echten Verständnis von Kausalität. Und die Abduktion (Bildung von Hypothesen für Beobachtungen, ‚beste Erklärung finden') ist tatsächlich eine große Schwäche aktueller KI. Kreativität, Intuition und das Generieren völlig neuer Konzepte sind hier angesiedelt.
> *Geschlossenes, selbstreferenzielles System und Autophagie*: Dieser Punkt ist besonders kritisch und weitsichtig. Die Tendenz von KI-Modellen, auf sich selbst zu referenzieren oder sich aus ihren eigenen generierten Daten zu ‚ernähren' (Autophagie), kann zu einer Reduzierung der Vielfalt und der Wirklichkeitserkennung führen. Wenn KI nur lernt, was KI schon generiert hat, können sich Verzerrungen, Fehler oder Vereinfachungen potenzieren. Das führt zu einer konservativen und manipulierbaren Informationsbasis.
> *Forderung nach einem technologiekritischen Ansatz*: Die Rufe nach einer Überwindung des behavioristischen und naiv-technizistischen Paradigmas sind absolut notwendig. Es ist wichtig, nicht nur zu fragen, was KI kann, sondern auch, wie sie es kann und welche Annahmen diesem ‚Können' zugrunde liegen. Eine kritische Metaphernanalyse der Anthropomorphismen (‚KI denkt', ‚KI versteht') ist entscheidend, um Missverständnisse zu vermeiden und eine sachliche Debatte zu ermöglichen.
> Mein Fazit: Ihr Text ist eine hervorragende Grundlage für eine differenzierte Auseinandersetzung mit den Fähigkeiten und Grenzen von KI. Er geht über die reine Leistungsbetrachtung hinaus und beleuchtet die philosophischen Implikationen ...“

Danke für die Blumen – aber ist die KI hier also doch in eine Paradoxie gelaufen? Sie spricht über sich selbst und spricht sich dabei gleichzeitig die Möglichkeit ab, dass sie überhaupt über sich selbst sprechen kann. Sie sagt ja gerade, dass sie nicht denken und verstehen kann usw. *Da capo*: Wie intelligent kann Künstliche Intelligenz also sein?

Literatur

Alemohammad, Sina, Casco-Rodriguez, Josue, Luzi, Lorenzo, Humayun, Ahmed Imtiaz, Babaei, Hossein, LeJeune, Daniel, Siahkoohi, Ali, & Baraniuk, Richard G. (2024). *Self-consuming generative models go MAD*. International Conference on Learning Representations (ICLR), May 7th, 2024 to May 11th 2024.

Gransche, Bruno, & Manzeschke, Arne. (2024). Das bewegliche Heer der Künstlichen Intelligenz. Ein Technomythos als Summe menschlicher Relationen. In: Michael Heinlein & Norbert Huchler (Hrsg.), *Künstliche Intelligenz, Mensch und Gesellschaft. Soziale Dynamiken und gesellschaftliche Folgen einer technologischen Innovation* (S. 75–108). Springer.

Hug, Theo. (Hrsg.) (1990), *Die soziale Wirklichkeit der Theorie. Beiträge zur Theorievermittlung und -aneignung in der Pädagogik*.

Hug, Theo. (2018), Visualizing Archivals and Discursive Structures. Towards Enhanced Perspectives Using the Example of the Ernst-von-Glasersfeld-Archive. In: András Benedek & Kristóf Nyíri (Hrsg.), *Visualizing Archivals and Discursive Structures.* Budapest University of Technology and Economics, Hungarian Academy of Sciences, S. 161–171.

Hug, Theo. (2024). Desiderata der AI-Literacy-Diskurse. In Theo Hug, Petra Missomelius & Heike Ortner (Hrsg.), *Künstliche Intelligenz im Diskurs. Interdizsiplinäre Perspektiven zur Gegenwart und Zukunft von KI-Anwendungen* (S. 129–148) innsbruck university press (IUP).

Hug, Theo, Schorner, Michael, & Mitterer, Josef. (Hrsg.) (2013): *Ernst-von-Glasersfeld-Archiv. Eröffnung - Inauguration*. innsbruck university press (IUP)

Lyotard, Jean-Francois. (1987). *Der Widerstreit*. Fink.

Mainzer, Klaus, & Kahle, Reinhard. (2022). *Grenzen der KI – theoretisch, praktisch, ethisch*. Springer.

McCarthy, John, Minsky, Marvin, Rochester, Nathaniel, & Shannon, Claude. (1955). A proposal for the Dartmouth Summer Research Project on Artificial Intelligence, August 31. https://www-formal.stanford.edu/jmc/history/dartmouth/dartmouth.html

Niedermair, Klaus. (1987). Wittgensteins Tractatus und die Selbstbezüglichkeit der Sprache. Peter Lang.

Niedermair, Klaus. (1990). Sprachspiel, Sprachsystem, Sprachideal. Über die Möglichkeit einer selbstreferentiellen Sinntheorie im Anschluß an Ludwig Wittgenstein. In Theo Hug (Hrsg.), *Die soziale Wirklichkeit der Theorie. Beiträge zur Theorievermittlung und -aneignung in der Pädagogik* (S. 237–249).

Niedermair, Klaus. (1998). Über Bilder, Metaphern und Mythen der Informationsgesellschaft. Unterwegs zu einer sozialwissenschaftlichen Technologiekritik. In Theo Hug (Hrsg.), *Technologiekritik und Medienpädagogik. Zur Theorie und Praxis kritisch-reflexiver Medienkommunikation* (S. 103–125). Baltmannsweiler: Schneider-Verlag Hohengehren.

Niedermair, Klaus. (2001). Metaphernanalyse. In Theo Hug (Hrsg.), *Wie kommt die Wissenschaft zu Wissen? Bd. 2* (S. 144–165). Baltmannsweiler: Schneider-Verlag Hohengehren.

Niedermair, Klaus. (2010). *Eine kleine Einführung in Wissenschaftstheorie und Methodologie*. Studia.

Searle, John. (1980). Minds, Brains, and Programs. *Behavioral and Brain Sciences*, 3, 417–457.

Touring, Alan M. (1950). Computing Machinery and Intelligence. *Mind*, LIX (236), 433–460.

Wittgenstein, Ludwig. (1921). Tractatus logico-philosophicus. Suhrkamp.