Hans Böckler Stiftung
Mitbestimmung · Forschung · Stipendien

WORKING PAPER FORSCHUNGSFÖRDERUNG

Nummer 304, September 2023

Künstliche Intelligenz, Large Language Models, ChatGPT und die Arbeitswelt der Zukunft

Michael Seemann

Auf einen Blick
Die rasante Entwicklung von Systemen Künstlicher Intelligenz wie ChatGPT, die inhaltlich und sprachlich überzeugende Texte generieren können, hat eine intensive Debatte ausgelöst. Es stellt sich die Frage, welche Auswirkungen solche Systeme auf die Prozesse und Arbeitsweisen z. B. in Wissens- und Kreativberufen haben werden. Diese Literaturstudie wertet den aktuellen Stand der Debatte aus. Sie führt in die technische Grundlage, die sogenannten „Large Language Models“, ein und untersucht abschließend, welche Auswirkungen auf die Arbeitswelt zu erwarten sind.

FORSCHUNGS-
FÖRDERUNG

ISSN 2509-2359

Inhalt

Abbildungen

Zusammenfassung

In dieser Studie wurden wissenschaftliche Paper, journalistische Artikel und aktuelle Bücher ausgewertet, um einen allgemeinen Überblick über die Technologien und Fähigkeiten von Large Language Models zu geben und ihre Auswirkungen auf die Arbeitswelt abzuschätzen.

Die Ergebnisse im Überblick:

- Large Language Models (LLMs) sind eine bahnbrechende, neue Technologie. Obwohl LLMs nur versuchen das jeweils nächste Wort eines Textes statistisch vorherzusagen, erlangen sie dadurch die Fähigkeit auf komplexe Konversationen zu reagieren, Anweisungen auszuführen, Denkaufgaben zu lösen und gut lesbare Texte zu schreiben. Noch sind sie für viele Zwecke ungeeignet und produzieren immer wieder Fehler, aber von einer rasanten Weiterentwicklung ist auszugehen.
- Die Frage, ob und was LLMs von dem, was sie ausgeben, „verstehen" ist fachlich umstritten und leidet an unzureichenden Metriken und nicht gesicherten Indikatoren. Für die spezifische Art des Erfassens von Bedeutung durch LLMs fehlt uns das passende Vokabular.
- Studien über Arbeitsplatzverlust und Produktivitätsgewinne durch LLMs sind in der Frühphase und leiden an bereits oft kritisierten methodischen Mängeln und sollten trotz ihrer Empirie als weitgehend spekulativ angesehen werden. Gewinnbringender scheint, über Strukturveränderungen, neue Überwachungsmöglichkeiten und übergeordnete Effekte nachzudenken, die als veränderte Umweltbedingungen der Kommunikation diese auf den Kopf stellen können.
- Der Einzug von LLMs in die Unternehmen wird die Machtverhältnisse in der Wirtschaft neu sortieren. Zum einen werden Unternehmen immer abhängiger von Dienstleistern für Künstliche Intelligenz (KI), zum anderen werden Unternehmen versuchen, mittels KI Arbeitnehmer*innen stärker unter Druck zu setzen.
- Der Einzug von LLMs in die Arbeitswelt wird auf zwei Wegen stattfinden: einerseits werden LLMs vom Management an verschiedenen Stellen zur Kostenreduktion eingesetzt werden. Zum anderen werden Mitarbeiter*innen bei ihrer täglichen Arbeit von sich aus auf LLMs – im Zweifel heimlich – zurückgreifen.
- Der Einsatz von LLMs wird unterschiedliche Berufe auf drei verschiedene Arten treffen: Manche Berufe werden verschwinden, oder zumindest existenziell bedroht sein (Disruption). Andere Berufe werden durch LLMs nur im Randbereich tangiert (Integration). Einige Berufe werden nicht verschwinden, aber sich komplett neu erfinden müssen (Transformation).

1. Einleitung

In dem Film „I, Robot“ von 2004 – eine lose Adaption von Isaac Asimovs berühmter Kurzgeschichtensammlung mit demselben Namen – fragt der von Will Smith gespielte Ermittler Del Spooner den unter Mordverdacht stehenden Roboter Sonny, ob ein Roboter „eine Symphonie komponieren könne. Oder kann ein Roboter eine leere Leinwand in ein Meisterwerk verwandeln?“ Der Roboter fragt schnippisch zurück: „Kannst du das?“, woraufhin Spooner ins Grübeln kommt.

Dieser Ausschnitt zirkulierte schon öfters als Internet-Meme (Matt 2018). Doch seit dem Aufkommen generativer KI-Systeme hat eine Variante des Memes die Runde gemacht, in dem Sonny auf die Frage einfach nur „ja“ antwortet. Das Grübeln Spooners ist seitdem nicht mehr dasselbe.

Die Science Fiction, aber auch unsere kollektive Zukunftserwartung hatte viele Szenarien zu künstlicher Intelligenz auf dem Zettel. Aber dass KI ausgerechnet im Kreativbereich so früh, so enorme Fortschritte machen würde, stand nicht auf der Liste. Die Produktivitätsgewinne aus der Automatisierung, so war die allgemeine Vorstellung, sollten uns von der Arbeit befreien, und uns nicht Kunst streitig machen. Oder wie es der Twitternutzer Karl Sharro ausdrückte:

> „Dass Menschen die harte Arbeit verrichten, während die Roboter Gedichte schreiben und malen, ist nicht die Zukunft, die ich wollte.“ (Sharro 2023)

Diese Zukunft ist nun aber da und bringt, wieder einmal, unsere ganzen Vorstellungen von Arbeit durcheinander. Es ist gar nicht so lange her, dass sich Journalist*innen um die Zukunft der Kraftwagenfahrer*innen sorgten (Sorge 2016). Nun müssen sie um ihre eigenen Stelle bangen (Ropek 2023), während Kraftwagenfahrer*innen händeringend gesucht werden (Wolter 2023).

Unter den vielen neuen generativen KI-Systemen sticht vor allem eines heraus: das sogenannte Large Language Model (LLM), das spätestens seit Herbst 2022 in Form des Chatbots ChatGPT international Furore macht. Mit ChatGPT kann man chatten, ihm Fragen stellen, oder es Aufgaben erledigen lassen und das Programm scheint nicht nur auf Anhieb selbst komplexe Anweisungen sofort zu begreifen, es liefert auch erstaunlich detaillierte und teils überraschend kreative Antworten ab.

Künstliche Intelligenz konnte auch in der Vergangenheit schon vieles: es konnte Katzen von Hunden unterscheiden, es konnte Schach und Go spielen, sogar Proteine falten (Jumper et al. 2021). Aber mit ChatGPT gelang etwas Neues: es kann uns in unserer Sprache antworten.

Seitdem ist ein regelrechter Hype ausgebrochen. Jeden Tag erscheinen neue Paper mit neuen Durchbrüchen in der darunter liegenden Technologie. Das Silicon Valley ist aufgeschreckt. Google soll nach dem Erfolg von ChatGPT gar eine Vollversammlung einberufen haben, änderte die eigene Strategie und veröffentlicht seitdem ein eigenes Modell nach dem anderen (Grant/Metz 2023). Dabei war es Google, die die darunter liegende Technologie erfunden hatte (Vaswani et al. 2017). Mit den überall entstehenden generativen KI-Systemen haben die Risiko-Investoren neue Lieblinge, die nach der Enttäuschung der nicht eingetretenen Web3-Revolution (Geuter 2022) sehr willkommen sind.

Doch was steckt hinter dem Hype? Was ist das für eine Technik hinter ChatGPT und LLMs, und was kann sie wirklich? Und vor allem: was bedeutet das für die Zukunft der Arbeit? Nicht nur für die Arbeitsplätze, sondern auch für unser Arbeiten im Allgemeinen?

Das sind die Fragen, die in dieser Literaturstudie beantwortet werden sollen. Die Studie gliedert sich in vier Hauptkapitel.

Im ersten Kapitel gehe ich der Frage nach, wie LLMs technisch funktionieren. Im zweiten Kapitel versuche ich anhand der wissenschaftlichen Debatte nachzuvollziehen, was die LLMs wirklich können, wo ihre Grenzen liegen. Ich gehe auch der komplizierten Frage nach, ob LLMs wirklich „verstehen“, was sie antworten.

Im dritten Kapitel steht schließlich die Arbeitswelt im Fokus. Dazu analysiere ich die politische Ökonomie der LLMs, werte die Studienlage zu erwartenden Arbeitsplatzverlusten aus und thematisiere einige der zu erwartende Strukturänderungen, die die LLMs im Arbeitsleben auslösen könnten. Zuletzt gebe ich einen Ausblick in Form von drei fiktiven Beispielen der Arbeitswelt in zehn Jahren. Dabei nehme ich an, dass unterschiedliche Berufe auch sehr unterschiedlich von dem Einzug von LLMs in das Arbeitsleben betroffen sein werden. Besprochen wird der Arbeitstag einer Übersetzerin, eines Altenpflegers und einer Universitätsdozentin.

Einige der Fragen, die wir hier stellen, sind spekulativ, weswegen ich mir die Freiheit genommen habe, auch jenseits gesicherter Erkenntnis und referenzierbarer Literatur eigene Gedanken einzubringen und diese als solche zu kennzeichnen. Das gilt vor allem, aber nicht nur für die fiktiven Beispiele am Ende.

Diese Arbeit ist bereits umfangreich, dennoch konnte ich nicht alle Aspekte dieser komplexen Thematik gemäß ihrer Relevanz berücksichtigen. Die Fragen hinsichtlich Biases in KI, die Fragen nach ethischen KI-Systemen, sowie Fragen nach ökologischen Problemstellungen und staatlicher Regulierung werden in dieser Arbeit zwar angesprochen, aber nicht weiter vertieft.

Hinsichtlich KI und Geschlechtergerechtigkeit gibt es aber bereits eine aussagekräftige Studie von Kathrin Ganz und Tanja Carstensen: „Vom Algorithmus diskriminiert? Zur Aushandlung von Gender in Diskursen über künstliche Intelligenz und Arbeit“, die ebenfalls im Rahmen der Hans Böckler Stiftung erschienen ist (Carstensen/Ganz 2023). Zudem empfiehlt sich für eine allgemeinere Einschätzung des Themas LLMs das Papier des Büros für Technikfolgenabschätzung beim Deutschen Bundestag (Albrecht 2023).

Ein letzter Hinweis: Ich spreche in erster Person, wenn es darum geht, eigene Entscheidungen zu erklären, wie in diesem Satz. Ich spreche von „wir“, wenn ich zusammen mit den Leser*innen dieser Studie – also Sie alle – gemeinsam einen Sachverhalt anschauen, reflektieren oder erinnern.

2. Was sind Large Language Models und wie funktionieren sie?

Künstliche Intelligenz im allgemeinen und Large Language Models im Besonderen sind komplizierte Technologien mit einer längeren Herkunftsgeschichte, deren Verständnis eine Menge Vorwissen erfordert. Um ihre Grenzen und Potenziale abschätzen zu können, ist ein Grundverständnis der Technik unabdingbar.

Deswegen steigen wir im ersten Kapitel tief in die Funktionsweise von Large Language Models ein. Ziel ist ein niedrigschwelliger Zugang zum Thema, jedoch gleichzeitig ein ausreichendes Verständnis aller Aspekte, die diese Form der KI auszeichnet und so erfolgreich macht.

Wir beginnen mit Begriffsdefinitionen und erklären dann, wie die Modelle das nächste Wort generieren. Ein kurzer Ausflug in die Geschichte des Deep Learning hilft dem Hintergrundverständnis. Erst dann kommen wir zu dem technologischen Sprung, der die neuen Modelle so erfolgreich macht: die sogenannte Transformer-Architektur. Im Anschluss schauen wir uns den Trainingsprozess solcher Modelle an, besprechen die aktuell wichtigsten Modelle und ihre Unterschiede und loten zuletzt die Entwicklungspotenziale der Technologie aus.

2.1 Begriffe zur Einführung

Künstliche Intelligenz (KI) ist ein Feld der Informatik, das fast so alt ist wie die Informatik selbst. In der KI geht es darum, Computer dazu zu bringen, auf bestimmte Arten zu agieren, die von Menschen als intelligent empfunden werden. Das schließt unter anderem die Lösung von komplexen Problemen, das selbstständige Lernen von neuen Fähigkeiten und auch die Beherrschung der menschlichen Sprache mit ein.

Künstliche Neuronale Netzwerke (KNN) sind die derzeit meistverwendete Technologie im Feld der KI. KNN bestehen aus künstlichen Neuronen und sind von den neuronalen Netzwerken im Gehirn von Menschen und Tieren inspiriert. KNN werden in einem Prozess namens „Deep Learning“ oder auch „maschinelles Lernen“ mit großen Datenmengen trainiert und erlangen dadurch Fähigkeiten, die schwer wären, durch normale Programmierung herzustellen; etwa das Erkennen von Objekten, Menschen oder Katzen, oder die Fähigkeit, Texte zu generieren, die Texten menschlichen Ursprungs ähneln.

Natural Language Processing (NLP) ist das Feld der KI, das sich dem maschinellen Analysieren, Transformieren und Generieren von natürlicher Sprache widmet.

Large Language Models (LLM) sind Künstliche Intelligenzen, die auf das Gebiet von NLP spezialisiert sind und aufgrund ihrer beachtlichen Fähigkeiten zur aktuell breit geführten Debatte um KI beigetragen haben. LLMs basieren auf KNN und sie stehen im Fokus dieser Literaturstudie.

Generative Pre-Trained Transformer (GPT) sind die derzeit populärsten LLM-Systeme. Die Firma OpenAI hat mit ihrem Chatbot ChatGPT und Modellen wie GPT-4 derzeit den größten Erfolg. Obwohl auch die meisten anderen LLMs technisch zu den GPTs gezählt werden können, verwendet vor allem OpenAI den Begriff für seine Systeme.

Tokens sind in ganze Zahlen umgewandelte Worte oder Wortbestandteile, wobei jedem Wort eine feststehende Zahl zugewiesen ist. Wenn LLMs trainiert werden, müssen die Trainingsdaten in Tokens umgewandelt werden. Wenn LLMs Sprache verarbeiten oder generieren, verarbeiten sie Tokens und generieren Tokens, die am Ende wieder in Worte umgewandelt werden.

Parameter sind die gewichteten Verbindungen zwischen den künstlichen Neuronen in KNN. In den Parametern liegen die Informationen gespeichert, mit denen eine KI, die auf KNN basiert, arbeitet. Die Anzahl der Parameter gibt eine ungefähre Vorstellung von der Größe und Komplexität und damit auch Leistungsfähigkeit einer KI.

Das **Kontext-Fenster** (Context Window) umfasst bei LLMs den Kontext eines aktuell zu generierenden Wortes. Da LLMs immer nur das nächste Wort vorhersagen, geschieht diese Vorhersage unter Einbezug aller vorher geschrieben Worte (Tokens), inklusive der Eingabe der Nutzer*innen. Das Kontext-Fenster fungiert somit wie der Arbeitsspeicher eines LLM.

OpenAI ist die Firma, die die derzeit erfolgreichsten und bekanntesten LLMs wie GPT-3.5 und GPT-4 über den Chatbot ChatGPT bereitstellt. Sie wurde 2015 als Non-Profit gegründet, um einen offenen und ethischen Ansatz der KI-Entwicklung zu verfolgen, aber agiert seit 2019 als gewinnorientiertes Startup, das mit Investorengeld Produkte entwickelt und seine wichtigsten Technologien geheim hält. Seit dieser Zeit ist es auch operativ und finanziell eng an Microsoft gebunden.

2.2 Es werde das nächste Wort

LLMs sagen immer nur das nächste Wort voraus. Das klingt trivial und ein bisschen so, wie die Wortvorschläge beim Nachrichten-Tippen auf dem

Smartphone. Der wesentliche Unterschied zu dieser recht einfachen Technologie besteht darin, dass das Smartphone für eine Wortvorhersage nur vom letzten geschriebenen Wort aus rät. LLMs nehmen dagegen die gesamte Sequenz an geschriebenen Worten als Ausgangspunkt für die Vorhersage.[1]

Es ist leicht, nach dem Wort „Ich“ ein „bin“ vorherzusagen. Aber wie wird das nächste Wort nach dem Satz, den sie gerade lesen, sein? Wie wird der der Absatz, oder der gesamte Text dieser Studie zu Ende gehen? Natürlich unter Berücksichtigung seiner gesamten bisherigen Struktur, seiner Argumente, dem Schreibstil in dem er verfasst ist, sowie den gesamten Kontext des zu behandelten Themas? An dieser Aufgabe kann man nur scheitern. Aber heute scheitern LLMs besser an dieser Aufgabe als viele Menschen.

Doch was heißt „besser“ in diesem Zusammenhang? Das qualitative Urteil, das an Sprachmodelle herangetragen wird, ist eines der Täuschung. Wenn ein LLM gut ist, meinen wir, dass ihre Resultate uns überzeugen könnten, von einem Menschen verfasst zu sein (Natale 2021). Dazu wurde das System mit Millionen von Menschen geschriebenen Texten gefüttert, die es statistisch durchmessen hat, sodass es anhand dieser Statistik die Wahrscheinlichkeit des nächsten Wortes in einem Satz, Absatz oder Text vorhersagen kann.

In diesem Prozess der statistischen Durchforstung endloser Textmengen hat die Maschine „gelernt“, wie Worte sich statistisch zueinander verhalten. Sprache ist vielseitig und komplex, doch in ihr gibt es auch eine ganze Menge Regelmäßigkeiten. Das LLM lernt z. B. schnell die Regel, dass auf ein Subjekt irgendwann ein Prädikat und dann irgendwann ein Objekt folgt. Das LLM lernt Grammatik, ohne, dass ihm jemand die Subjekt-Prädikat-Objekt-Regel explizit einprogrammieren müsste. Syntaktik und Grammatik sind statistisch vergleichsweise leicht abzuleiten; sie sind so einfach, dass wir sie sogar sie in expliziten Regeln aufschreiben konnten.

Es gibt aber auch eine Menge Regeln bzw. Regelmäßigkeiten in der Sprache, die wir bislang noch gar nicht formell festgehalten haben, weil sie so komplex sind. Nehmen wir den Satz: „Die Wirtschaft spaziert Aschenbecher in der Nadel.“ Das ist ein grammatikalisch wohlgeformter Satz, aber er macht keinen Sinn. Wir können zwar erklären, warum dieser Satz keinen Sinn ergibt, aber wir haben keine allgemeinen Regeln dafür, wie man sinnhafte Sätze formt.

1 Dieser Unterschied wird vielleicht nicht mehr lange existieren. Apple hat bereits für sein aktuelles Smartphone-Betriebssystem angekündigt, die Wortvorschläge und Autokorrekturen in Textnachrichten tatsächlich mithilfe eines GPT zu generieren (Mayo 2023).

Der überraschende Erfolg der LLMs basiert darauf, dass sie auch semantisch korrekte Sätze zu formen imstande sind. Sprachmodelle kamen bislang immer dort an ihre Grenze, wo die Satzkonstruktion ein gewisses Verständnis des Inhalts erfordert. Etwa bei hierarchischen Satzkonstruktionen wie: „Die Schlüssel zum alten, moderigen Schuppen lagen auf dem Tisch“. Es erfordert ein Verständnis des Inhaltes des Satzes, um das Verb „lagen“ (Mehrzahl, Vergangenheitsform) richtig zu bilden, weil es sich auf das weit zurückliegende „Schlüssel“ bezieht (Mahowald et al. 2023).

Es ist rechnerisch leicht, eine Wahrscheinlich für das Wort nach einem anderen Wort zu berechnen. Man nennt solche Wortpaare „2Grams“. Schon deutlich schwieriger wird es, wenn man ein 3Gram berechnen will, d. h. von zwei Worten aus das dritte zu berechnen. Auf einmal hat man zwei abhängige Variablen, die es zu berücksichtigen gilt und mit jedem zusätzlichen Wort steigen die nötigen Rechenoperationen exponentiell.

Nun könnte man sich vorstellen, aus den gesamten Trainingsdaten die vorkommenden n-grams (n steht für die Anzahl der verkoppelten Worte) zu bilden und zu speichern. Man erhielte eine verlustfreie Kopie der Trainingsdaten. Jedoch würde so eine Prozedur schnell die Rechenkapazitäten aller verfügbaren Computer der Welt sprengen.

Was man stattdessen macht, ist das, was man immer macht, wenn die Realität für die Verarbeitung zu komplex ist: Man macht ein Modell. Ein Modell ist immer eine Annäherung an die Realität, die nicht perfekt, aber für bestimmte Zwecke gut genug ist. LLMs können also wortwörtlich als ein Modell der menschlichen Sprache verstanden werden, so wie eine Modelleisenbahn ein Modell einer wirklichen Eisenbahn ist. So wie eine Modelleisenbahn sich bemüht, alle möglichen Details der Eisenbahn abzubilden, so versucht ein LLM alle möglichen Details von Sprache abzubilden, aber ist dabei, wie die Modelleisenbahn, eben nur so gut, wie es die Technik gerade zulässt.

Das hat einige Implikationen. Wären die Trainingsdaten in n-grams gespeichert, wäre ein LLM in der Lage, alle Fakten, Quellen und Zitate der Trainingsdaten Buchstabe für Buchstabe wiederzugeben. Weil ein LLM aber nur ein Modell der Sprache ist, klappt das nur manchmal.

In den Daten klaffen Lücken und das Modell ist besonders kompetent darin, diese Lücken so zu füllen, dass es so aussieht, als seien da gar keine Lücken. So kommt es vor, dass es enorm selbstsicher formulierte Sätze ausgibt, deren Fakten ausgedacht sind und sie beim genauen Hinschauen teils überhaupt keinen Sinn ergeben. Man spricht dann z. B. davon, dass LLMs „halluzinieren“. Was sie aber eigentlich tun, ist die Lücken zu füllen mit Worten, die statistisch plausibel hineinpassen.

Ted Chiang, Autor beim Magazin The New Yorker, brachte diese Modellhaftigkeit der Sprachmodelle gut auf den Punkt, indem er sie mit

JPEGs verglich. (Chiang 2023) JPEG ist so etwas wie das Standardformat für Fotos im Internet und ist bekannt für seine enorm effektive, aber verlustreiche Kompression. Speichert man seine Fotos im JPEG-Format, lassen sich eine Menge Daten sparen, doch schaut man genauer auf die Details in den Fotos, fallen einem die verschwommenen Fragmente an den Rändern von Konturen ins Auge. ChatGPT sei ein verschwommenes JPEG des Internets, so der Autor.

JPEGs sind für viele Zwecke ungenügend, vor allem Profis greifen lieber auf verlustfreie Kompressionsformate zurück. Dennoch haben JPEGs einen enormen Nutzen und das gilt offenbar auch für LLMs, zumindest wenn man ihre Stärken und Schwächen kennt. Wie immer gilt der Ausspruch von George Box: „Alle Modelle sind falsch, aber manche sind nützlich." (Box 1979, S. 202).

2.3 Kontext: Die Deep-Learning-Revolution vor zehn Jahren

Das Training von LLMs ist sehr gradlinig. Man gibt der Maschine einen Teil von einem Text und bittet sie, das nächste Wort zu ergänzen. Dann vergleicht man das geratene Wort mit dem tatsächlich im Text folgenden Wort, errechnet den Fehlerwert und gibt das Ergebnis an das System zurück, das die Information dazu verwendet, seine Vorhersagefähigkeiten zu verbessern.

Am Anfang wird die Maschine noch irgendwelche zufälligen Wörter vorschlagen, die nicht mal ansatzweise Sinn ergeben. Etwa „Ich Tasse" oder „Mensch rot". Die dadurch ausgelösten negativen Feedbacksignale helfen aber jedes Mal ein kleines Stück, das System zu verbessern. Je öfter das System diesen Prozess durchgemacht hat – wie sprechen hier von Hunderte Milliarden Mal – desto besser wird es im Raten.

Das dahintersteckende Verfahren ist sehr viel älter als die aktuellen Sprachmodelle und nennt sich „Deep Learning". Beim Deep Learning geht es darum, mittels großer Datenmengen ein künstliches neuronales Netz zu trainieren. KNN sind von natürlichen neuronalen Netzen wie die Gehirnstrukturen von Menschen und Tieren inspiriert.[2]

Die ersten künstlichen Neuronen im sogenannten „Perceptron" von 1958 waren tatsächlich noch Hardware-Relais, die mit Drähten verbunden waren (Loiseau 2019). Experimente mit KKN auf Softwarebasis wurden seit den 1970er-Jahren immer wieder gemacht, doch bis auf wenige Ein-

2 Zu den tatsächlichen Unterschieden von künstlichen und natürlichen Neuronen siehe Nagyfi 2018.

satzzwecke z. B. in der E-Mail-Spamerkennung hatte der Ansatz nur wenig Relevanz.

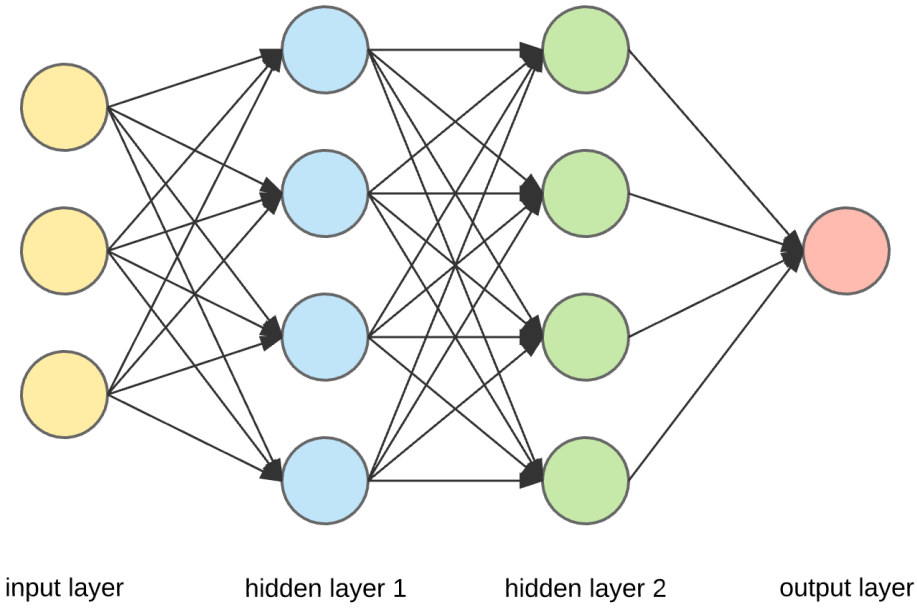
Erst im Jahr 2012 gelang der Durchbruch. Bis dahin gab es unterschiedliche Ansätze, Künstliche Intelligenz voranzubringen, etwa symbolische Systeme oder Expertensysteme. Bei diesem Ansatz versuchen Menschen die zu lösenden Aufgaben in Form von Regeln zu definieren und diese Regeln als Code im KI-System zu implementieren. Das war noch bis 2010er-Jahre hinein einer der wichtigsten KI-Ansätze.

Das änderte sich mit der „ImageNet Challenge“ von 2012, einem Wettbewerb in künstlicher Bilderkennung. Damals hatte sich das Team unter der Leitung von Geoffrey Hinton mit einem „Deep Neural Net“ (ein anderer Name für KNN) namens AxelNet mit großem Abstand zu allen anderen Bewerbern durchgesetzt. Seitdem dominieren KKN das gesamte Feld von KI und selbstlernender Systeme, nicht nur in der Bilderkennung.

Die künstlichen Neuronen eines KNN sind in mehreren Ebenen („layers“) angeordnet, und zwischen den Ebenen durch künstliche Synapsen miteinander verbunden. Die erste Ebene ist die Input-Ebene, also die Reihe von Neuronen, auf die die eingehenden Daten treffen. Am Ende steht die Output-Ebene, die das Ergebnis der Berechnung liefern soll: ist ein Hund oder Katze auf dem Bild, oder was ist das nächste Wort?

Zwischen Input und Output sind mehrere sogenannte „versteckte Ebenen“ („hidden layers“) geschaltet, die die eigentliche Informationsprozessierung handhaben. Grob gesagt kann ein künstliches Neuronales Netz mit mehr versteckten Ebenen komplexere Aufgaben bewältigen und je mehr versteckte Ebenen es gibt, desto aufwendiger sind die Durchläufe durch das Netz.

Abbildung 1: Vereinfachtes Künstliches Neuronales Netzwerk



Quelle: Dertat 2017

Treffen Daten auf die Input-Ebene, entscheiden die einzelnen Neuronen jeweils anhand einer integrierten Funktion, ob und wie stark sie das Signal an die dahinterliegende versteckte Ebene weiterreichen sollen. Die Neuronen der versteckten Ebene haben ebenfalls eine Funktion, um zu entscheiden, wie die eintreffenden Signale der Input-Ebene gedeutet werden sollen und geben davon abhängig ihrerseits ein Signal an die Neuronen der nächsten versteckten Ebene weiter.

Und so wandern die vom Input ausgelösten Signale von Ebene zu Ebene, wobei jedes Neuron von sich aus entscheidet, welchen Output es auf welchen Input hin weitergibt. Je weiter sich die Signale durch die versteckten Ebenen arbeiten, desto abstraktere Informationen werden für gewöhnlich verarbeitet.

Hier ein vereinfachtes Beispiel für eine Bilderkennung: Die Input-Ebene bekommt ein Bild und teilt es in Bereiche für ihre Neuronen auf. Die erste versteckte Ebene identifiziert darin Kontraste, die zweite Ebene interpretiert die Kontraste und erkennt Formen, die dritte Ebene fügt die Formen zu einer Gesamtkomposition zusammen und die Output-Ebene gibt Wahrscheinlichkeitswerte aus, welchen Objekten die Gesamtkompo-

sition ähnlich sieht. Jeder Input durchläuft das gesamte neuronale Netzwerk, bis es an der Output-Ebene zu einer Entscheidung kommt.[3]

Befindet sich das neuronale Netz im Trainingsprozess, wird ihm nach jedem Durchlauf zurückgespiegelt, ob es richtig lag. Dieses Feedback wird dann wiederum im ganzen Netzwerk verarbeitet, nur läuft es diesmal rückwärts. Dabei errechnet eine sogenannte „Verlust-Funktion“ (Loss Function), wie weit das Netzwerk vom richtigen Ergebnis entfernt lag.

Bei einer falschen Bilderkennung schaut die Output-Ebene, welche Verbindungen (Parameter), der letzten versteckten Ebene dazu beitrugen, die falsche Entscheidung zu treffen und reduziert ihre Relevanz durch Anpassung der Gewichtungen. Dasselbe tut die versteckte Ebene mit ihrem Vorgänger, und diese wiederum mit ihrem Vorgänger und so weiter, bis hin zur Input-Ebene.

Dieses automatische Einarbeiten von Feedback durch das ganze Netz nennt man „Backpropagation“ und ist zentraler Bestandteil aller heutigen KNN, inklusive der LLMs. Dabei steht die Minimierung der Verlust-Funktion im Zentrum. Wenn der errechnete Verlustwert nicht mehr sinkt, ist das Netzwerk im Rahmen der gegebenen Möglichkeiten fertig trainiert.

Dass der Durchbruch dieser Technologie erst im Jahr 2012 erfolgte, obwohl daran seit den 1970er-Jahren geforscht wird, lag an zwei Faktoren: Zum einen können derart signifikante Ergebnisse nur mit einer enormen Menge von Trainingsdaten erzielt werden, so vielen Daten, wie sie vor der massenweisen Verwendung des Internets den Wenigsten zur Verfügung standen.

Der zweite Faktor war, dass das Training eines neuronalen Netzwerks ab einer bestimmten Komplexität enorme Rechenkapazitäten erfordert. Durch die Popularisierung von Videospielen standen ab den 2010er-Jahren leistungsfähige Grafikprozessoren (GPUs) zur Verfügung, die bestimmte Berechnungen schneller als herkömmliche Prozessoren und dazu noch parallel durchführen konnten.

Durch diese beiden Faktoren konnten immer komplexere Modelle mit immer mehr versteckten Ebenen auf noch mehr Daten trainiert werden, und dieses Skalieren von Ebenen, Parametern und Trainingsdaten führt bis heute zu immer neuen Fähigkeiten von KNN.

3 Das ist, wie gesagt, ein vereinfachtes Beispiel zur Veranschaulichung der ungefähren Verarbeitungsmechanik. In Wirklichkeit ist sie wesentlich komplexer, und Wissenschaftler*innen haben große Schwierigkeiten, bestimmte Informationsverarbeitungsschritte einem Neuron oder eine Ebene zuzuordnen. Tatsächlich ist die Verarbeitungsmechanik ein eigenes Forschungsfeld (vgl. Saeed/Omlin 2023).

2.4 Der Trainingsprozess: Einbettung im latenten Raum

Auch die heutigen LLMs, basierend auf dem Transformermodell (dazu gleich mehr), müssen trainiert werden. Zuvor müssen allerdings die Trainingsdaten bereitgestellt und bearbeitet werden. LLMs sind mit Billionen von Wörtern trainiert und die müssen erstmal gesammelt werden. Dabei spielt das Internet eine wichtige Rolle. Selbst riesige Textsammlungen wie die Wikipedia oder die Summe aller digitalisierten Bücher machen nur einen Bruchteil der Trainingsdaten aus. Der größte Teil der Texte kommt aus rudimentär gesäuberten Datensammlungen, die im Grunde aus Millionen beliebig zusammengesuchten Websites bestehen.[4]

In einem zweiten Schritt müssen diese Trainingsdaten für das Modell in Tokens umgewandelt werden. Dafür gibt es sogenannte Tokenizer, also Programme, die jedem Wort oder Wortbestandteil, eine ganze Zahl zuordnen. Mit jedem Token wird dann ein vieldimensionaler Vektor verknüpft – ein sogenanntes „Embedding" – eine „Einbettung" in den Kontext aller anderen Tokens. Im Embedding sind alle Beziehungen eines Tokens zu allen anderen Tokens gespeichert. Zu Beginn des Trainings ist dieser Vektor allerdings noch mit rein zufälligen Werten belegt.

Wenn das Modell im Laufe des Trainings dann für jede Abfolge von Tokens den jeweils nächsten Token rät, wird die Vorhersage mit dem Ergebnis des tatsächlich nächsten Tokens in den Trainingsdaten verglichen und die Abweichung der Verlust-Funktion per Back-Propagation durch das Netzwerk zurückgefüttert. Im Zuge dieses Lernprozesses werden nicht nur die Verbindungen neu gewichtet, sondern es werden auch mit jedem Schritt die Embeddings der Tokens aktualisiert. Auf diese Art bilden sich im Zuge des Trainings die Beziehungen der Tokens zueinander immer deutlicher heraus.

Am Ende dieses langen Prozesses ist in den Embeddings die Komplexität von sprachlichen Äußerungen nicht nur auf Wort- oder Satzebene, sondern auch auf Konzept- und Ideen-Ebene gespeichert. Es entwickelt sich eine 1000-dimensionale Landkarte (bei GPT-3.5 sind es 12.288 Dimensionen) der Sprache. In dieser Landkarte ist hinterlegt, wie sich „Rot" zu „Vorhang" verhält, „Liebe" zu „Haus" und „Zitronensäurezyklus" zu „Salat". Das Modell kennt diese Dinge nicht aus eigener Anschauung, aber es hat aus den Millionen Texten erfahren, in welche vielfältigen Verhältnisse wir diese Begriffe zueinander setzen.

4 Um einen Eindruck zu bekommen, wie diese Trainingsdaten aufgebaut sind, hat die Washington Post ein populäres Trainingsdaten-Set in einer Infografik aufbereitet (Schaul/Chen/Tiku 2023).

Diese Landkarte wird auch als „latent space“, als latenter Raum bezeichnet. Im latenten Raum liegen semantisch ähnliche Wörter nahe beieinander und semantisch unähnliche sind weiter entfernt. Ein vereinfachtes Beispiel: Zieht man vom Embedding „König“ das Embedding „Mann“ ab und addiert das Embedding „Frau“, landet man im Latenten Raum beim Embedding „Königin“ (Mikolov et al. 2013).[5]

LLMs sind kompetent, auf dieser Landkarte zu navigieren. Gibt man z. B. GPT-3.5 einen Textanfang, dann ist das, als hätte man dem Modell einen Pfad auf dieser Landkarte vorgezeichnet und es am Endpunkt des Pfades abgesetzt mit der Aufgabe, ihn selbstständig weiterzugehen. Das ist eine anspruchsvolle Aufgabe, gibt es doch auf jeder der 12.288 Dimensionen Nähen und Fernen zu anderen Tokens (z. B. assoziative Nähen und Fernen, funktionale Nähen und Fernen, phonetische Nähen und Fernen etc.).

Dabei sind zwar alle Tokens des bereits zurückgelegten Weges mit in Betracht zu ziehen, doch der Aufmerksamkeitsmechanismus hat wichtige Wegmarken nochmal gesondert gekennzeichnet, um Orientierung zu geben. Wie ein Pfadfinder sucht GPT-3.5 nun nach möglichst ausgetretenen Pfaden, die mit dem Herkunftspfad und den Orientierungsmarken in Übereinstimmung zu bringen sind.

Eine interessante Besonderheit bei LLMs ist, dass man Einfluss nehmen kann, wie ausgetreten die Pfade sein sollen, die das Modell aussucht. Die einfachste Idee wäre, tatsächlich immer das wahrscheinlichste Wort zu nehmen und auszuspucken. Es hat sich jedoch gezeigt, dass die Texte dadurch oft sehr starr und wenig interessant werden und dass sie schnell dazu tendieren, sich zu wiederholen. Deswegen kann man über die „Temperatur“ die „Wildheit“ des Modells einstellen.

Temperatur ist ein Wert, der angibt wie oft das Modell nicht das wahrscheinlichste, sondern auch mal das zweit- oder drittwahrscheinlichste Wort als Vorhersagen verwenden soll. Bei einer Temperatur von 0,1 wird das Modell sehr konsistente, aber langweilige Texte produzieren, bei einer Temperatur von 0,9 kommt kaum mehr verständliches Gerede bei rum. Meist wird deswegen mit einer Temperatur um die 0,7 gearbeitet (Wolfram 2023).

5 Das Beispiel ist aus dem Vektorraum eines Sprachmodells aus dem Jahr 2013, also vor der Erfindung der Transformer-Modelle, es sollte aber auch für aktuelle Systeme eine ungefähre Gültigkeit haben.

2.5 Der Aufstieg der Transformer-Modelle

Seit 2012 haben sich viele unterschiedliche Architekturen für KNNs durchgesetzt. Multilayer Perceptrons (MLP), Convolutional Neural Networks (CNN) und Recurrent Neural Networks (RNN) waren vor den Transformermodellen die populärsten Architekturen. Sie sind heute überall zu finden. In Fotosoftware, Suchalgorithmen, oder auch in der Industrie in den unterschiedlichsten Anwendungen.

Der technologische Durchbruch, der die aktuell erfolgreichen generative KIs wie ChatGPT, aber auch Bildgeneratoren wie Midjourney und Stable Diffusion ermöglicht hat, wurde 2017 durch Forscher*innen bei Google in einem Paper mit dem Titel „Attention Is All You Need“ (Vaswani et al. 2017) beschrieben. In dem Aufsatz wird das sogenannte Transformer-Modell beschrieben. Das ist eine Architektur für KNN, die jeder versteckten Ebene (in diesem Fall heißt sie „Feed-Forward-Ebene“) eine sogenannte Aufmerksamkeits-Ebene zur Seite stellt (Nyandwi 2023). Die soll ihr helfen, den relevanten Kontext der aktuellen Aufgabe besser im Blick zu behalten, indem er den in den Embeddings zusätzlich vermerkt wird.

Aufmerksamkeit ist deswegen wichtig, weil z. B. beim Generieren des nächsten Tokens zwar der ganze Kontext (alles vorher Geschriebene) mit in Betracht gezogen werden muss, aber eben nicht alles gleich stark (Serrano 2023). Um den Satz „Die Schlüssel liegen dort, wo ich sie hingelegt habe.“ zu schreiben, muss das System z. B. beim Generieren des Wortes „sie“ wissen, dass es sich auf „Schlüssel“ bezieht. Das Wort Schlüssel ist in diesem Moment des Generierens von „sie“ also wichtiger als die anderen Worte des Satzes.

Die Aufmerksamkeits-Ebene assistiert der versteckten Ebene, indem sie den jeweiligen Kontext des zu bearbeitenden Tokens nach Relevanz sortiert und entsprechend gewichtet. Für jeden Token im Kontext-Fenster wird die Relevanz jedes anderen Tokens berechnet und diese in seinen Embeddings vermerkt. Das hilft nicht nur dabei, grammatikalische Konsistenz zu erhalten. Die Tokens (und damit die multidimensionalen Embeddings) durchwandern das ganze Netzwerk von Aufmerksamkeits- und Feed-Forward-Ebenen und werden von jeder einzelnen der Aufmerksamkeitsebenen mit neuen Kontexten angereichert, die als neue Dimensionen im jeweiligen Embedding vermerkt werden.

Heutige Modelle haben sehr viele von diesen Ebenen, bei GPT-3.5 sind es über 90. In den tieferliegenden Ebenen, dort wo abstraktere Aspekte prozessiert werden, hilft der Attention-Mechanismus dem Modell unter anderem narrative oder konzeptionelle Kohärenz eines Textes zu gewährleisten (Lee/Trott 2023).

Wenn man z. B. einen Text mit der Hochzeit von Alice und Bob anfängt, „versteht“ das System, dass diese Hochzeit und ihre Protagonisten wichtig für die Fortführung des Textes bleiben und schreibt ihn fort, ohne dabei den thematischen Fokus zu verlieren. Oder wenn man das Modell bittet, eine komplexe Denkaufgabe zu lösen, hilft der Attention-Mechanismus, sich auf die wesentlichen Bestandteile der Antwort zu konzentrieren.

Es stellt sich heraus, dass fokussierte Aufmerksamkeit auf jeder Abstraktionsebene auf unterschiedliche Weise hilfreich ist. Am Ende, wenn das System die Entscheidung darüber trifft, welches nächste Wort nun ausgegeben wird, sind alle bisherigen Tokens des Kontext-Fensters mit abertausenden zusätzlichen kontextbezogenen Dimensionen angereichert, die mit in die Berechnung einbezogen werden.

Die Transformer-Architektur ermöglicht es dem System zudem, die Aufmerksamkeits-Gewichtungen parallel für viele Worte gleichzeitig zu berechnen. Das spart Zeit und erklärt, warum Grafikprozessoren (GPUs) mit vielen Prozessorkernen eine wichtige Ressource für die Entwicklung von aktuellen LLMs darstellen.

Eine weitere Eigenschaft von Transformer-Modellen ist, dass sie nach dem Training noch verfeinert werden können. Nach dem initialen Training erhält man ein sogenanntes „pre-trained model“ oder auch „foundational model“ genannt. Die durch das Pre-Training erworbene Kompetenz im Interpretieren von Sprache kann dann für die Feinabstimmung des Systems genutzt werden. So kann es durch ein sogenanntes „reinforcement learning by human feedback“ auf bestimmte Aufgaben optimiert werden, beispielsweise für Übersetzungsaufgaben, Textanalyse, Recherche oder den Einsatz als Chatbot wie ChatGPT.

Bei OpenAI z. B. geschieht das Reinforcement Learning in zwei Schritten: speziell geschulte Leute werden beschäftigt, um Beispiel-Prompts und dazugehörige „gute“ Antworten zu erstellen, mit denen das System weitertrainiert wird. Dabei werden vergleichsweise wenig Trainingsdaten verwendet, diese sind aber qualitativ hochwertig und werden im Training stärker gewichtet. In einem zweiten Schritt lässt man das Modell nach diesen Vorbildern selbst mehrere Antworten auf einen Prompt generieren und lässt Menschen die beste der Antworten auswählen (Karpathy 2023).

Josh Dzieza bringt die Rolle des Finetuning in einem Artikel für die The Verge auf den Punkt: „Anders ausgedrückt, scheint ChatGPT so menschlich zu sein, weil es von einer KI trainiert wurde, die Menschen nachahmte, die wiederum eine KI bewerteten, die Menschen imitierte, die so taten, als wären sie eine bessere Version der KI, die auf menschlichen Texten trainiert wurde.“ (Dzieza 2023).

2.6 Die aktuell wichtigsten Modelle

Es gibt mittlerweile eine ganze Flut an Sprachmodellen und beinahe jeden Tag kommen neue hinzu. Es ist schwierig den Überblick zu behalten, deswegen stellen wir hier nur eine kleine Auswahl der aktuell meist diskutierten LLMs vor.

Open AI hatte Anfang dieses Jahres mit GPT-4 den Nachfolger von GPT-3.5 veröffentlicht. GPT-3.5 ist das Modell, das für die meisten Menschen in Form von ChatGPT zur Verfügung steht. Nur zahlender Nutzer*innen bekommen aktuell auch Zugang zu GPT-4.

GPT-4 ist verglichen mit seinem Vorgänger ein vielen Dingen deutlich überlegenes Modell. Während GPT-3 mit 175 Milliarden Parametern trainiert wurde, gibt OpenAI für GPT-4 keine konkreten technischen Daten heraus: Gerüchteweise soll es sich um ein Multimodell oder Mixture of Experts (MoE) handeln, bestehend aus 16 auf Themengebiete und Fähigkeiten spezialisierten Foundation-Modellen mit jeweils 110 Milliarden Parametern, die je nach Fragestellung die Antworten geben (Barr 2023).

Eine weitere Besonderheit von GPT-4 ist, dass es „multimodal“ ist. Es wurde nicht nur mit Texten, sondern auch mit Bildern trainiert, was es theoretisch in die Lage versetzt, Bilder zu erkennen und evtl. auch produzieren. Diese Features sind aber bislang noch nicht freigeschaltet. Das einzig Offizielle, was man zu GPT-4 findet, ist ein Paper darüber, wie gut es in Tests abschneidet (OpenAI 2023a). Eine Besonderheit der OpenAI-Modelle ist die Integration von funktionserweiternden Plugins. Besonders hervorzuheben ist die Fähigkeit, zu browsen und Code zu interpretieren und auszuführen.

Neben OpenAI-Modellen wird viel über Googles neue Modelle mit dem Namen PaLM 2 und dem zugehörigen Chatbot Bard gesprochen. Der Vorgänger hatte bislang wenig beeindruckt, und auch die neuen Modelle scheinen nicht an die Qualität der OpenAI-Modelle heranzukommen (Pierce 2023; Owen 2023). PaLM 2 gibt es in vier verschiedenen Größen, von Gecko, das sogar auf Mobiltelefonen laufen soll, bis Unicorn, dem leistungsstärksten Modell. Auch Google hält sich bei der Veröffentlichung der Parameteranzahl bedeckt, aber es wird vermutet, dass die PaLM 2-Modellle mit 14,5 und bis zu 100 Milliarden Parametern trainiert wurden (Sha 2023).

PaLM 2 bzw. Bard wird bereits in viele Google-Produkte wie Gmail und Google Docs integriert und wird allein deswegen eine große Rolle spielen. Ein erst angekündigtes Modell namens Gemini soll 2024 veröffentlicht werden, von dem technisch aber viel erwartet wird.

Auch Meta Platforms (ehemals Facebook) hat eigene Modelle unter dem Namen LLaMA veröffentlicht (Meta 2023a; Touvron et al. 2023).

Auch wenn die erste Basisversion mit 65 Milliarden Parametern nicht beeindruckend war, war eine Besonderheit, dass Meta nicht nur den vollen Zugriff auf die Software, sondern auch auf die Parameter erlaubt. Nach einem aufwendigen Registrierungsprozess und dem Zustimmen der Lizenzvereinbarungen konnten sich Interessierte das gesamte Modell zur lokalen Installation herunterladen.

Es dauerte aber nicht lang und es tauchte als BitTorrent-Datei frei verfügbar im Internet auf (Vincent 2023). Damit war es das erste größere Open Source Basismodell und hat wiederum eine Kette von neuen Entwicklungen losgetreten.

Mittlerweile ist der Nachfolger LLaMA 2 erschienen, und zwar direkt unter Open-Source-Lizenz, was der Open-Source-Szene zusätzliche Schubkraft geben wird. LLama 2 hat bis zu 70 Milliarden Parameter und scheint an die Leistungsfähigkeit von GPT-3.5 heranzureichen (Meta 2023b).

Chinas wichtigstes Sprachmodell kommt von dem Suchmaschinenkonzern Baidu und soll nach eigenen Angaben besser sein als ChatGPT und nahe dran an GPT-4 (Che/Wang 2023). Der Chatbot heißt Ernie und spricht nur Chinesisch. Auch hier fehlen Angaben zu Architektur, Trainingsdaten und Parameteranzahl.

Ein Überraschungserfolg scheint das Modell Claude der Firma Anthrophic zu sein (Anthropic 2023). Die Firma wurde von einigen Aussteigern von OpenAI gegründet (Roose 2023a) und schon kurz nach der Veröffentlichung der ersten Version von Claude im März 2023 berichteten Nutzer*innen, dass das Modell fast an GPT-4 heranreicht.

Claude-V1 soll mit 52 Milliarden Parametern trainiert worden sein. Mit Claude 2 ist kurz vor Fertigstellung der Nachfolger erschienen, der in vielerlei Hinsicht sehr viel leistungsstärker sein soll, als sein Vorgänger (Hahn 2023). Unter anderem ist das Kontextfenster 100.000 Tokens lang, deutlich mehr als GPT-4 mit derzeit bis zu 16.000 Tokens (Saquib 2023). Mit Claude 2 lassen sich kurze Bücher oder lange Berichte übersetzen, zusammenfassen oder weiter befragen.

Die Frage, wie man die Leistungsfähigkeit von Sprachmodellen miteinander vergleicht, ist ein ungelöstes Problem. Es gibt einige standardisierte Tests und Benchmarks, die helfen, etwas Orientierung zu finden, aber deren Aussagekraft ist zweifelhaft. Zum einen gibt es immer die Möglichkeit, Sprachmodelle auf das Bestehen dieser Benchmarks hin zu optimieren, was nicht gleichbedeutend mit ihrer Nützlichkeit im Alltagsgebrauch ist. Zum anderen zeichnen sich gute Sprachmodelle gerade dadurch aus, dass sie durch Kreativität und Lösungskompetenz in Nicht-Standard-Situationen überraschen.

Die „Large Model Systems Organization“ (LMSYS Org) hat den bislang überzeugendsten Ansatz, um die Leistungsfähigkeit von LLMs vergleichbar zu machen. Auf ihrer Website (LMSYS Org 2023a) kann man über einen Prompt mit zwei unterschiedlichen LLMs simultan chatten, ohne zu wissen, welche Modelle man vor sich hat. Beide LLMs antworten parallel auf den gestellten Prompt und wenn man sich sicher ist, welches der Modelle besser antwortet, kann man entsprechendes Feedback geben. Dann wird angezeigt, mit welchen LLMs man geredet hat, und die Wertung fließt in ein sogenanntes ELO-Rating ein, das die unterschiedlichen LLMs dann entsprechend rankt (LMSYS Org 2023b).

Das Ranking wird zum Zeitpunkt des Schreibens von GPT-4 angeführt, gefolgt von Claude-1 und der etwas abgespeckten Version Claude V1 Instant. Erst dann kommt Claude -2, GPT-3.5 Turbo und darunter finden sich derzeit viele kleinere Unternehmen und Open Source Projekte. Googles PaLM 2 findet sich auf Platz 11 und Metas LLaMA 2 auf Platz 13. Es sei anzumerken, dass bei diesem Ranking einige wichtige Projekte (noch) nicht in der Wertung sind. Sei es, weil sie noch zu neu sind, oder sei es, weil sie keine API-Schnittstellen anbieten, über die die LLMs an die Website angebunden sein müssen.

2.7 Entwicklungspotenziale

Die Entwicklung von generativen KI-Systemen ist derzeit noch eher eine Kunst als eine Wissenschaft. Die Forschung basiert nach wie vor viel auf Trial-and-Error-Verfahren und nur zum Teil auf (meist nachträglicher) Theoriebildung. Gemacht wird, was funktioniert und mit der Zeit führt das angehäufte Erfahrungswissen zu Technologiesprüngen (Wolfram 2023).

Ohne der Frage vorzugreifen, wie die Entwicklung bei LLMs weitergeht (dazu mehr im nächsten Kapitel), lassen sich aus der rein technologischen Sicht bereits kurz- und mittelfristige Entwicklungsspielräume und Potenziale für die existierenden Architekturen identifizieren:

Seit 2020 werden die sogenannten „scaling laws“ diskutiert (Kaplan et al. 2020). Dabei geht es um die Frage, in welchem Verhältnis Trainingsdaten, Parameteranzahl und eingesetzte Computerressourcen zueinanderstehen müssen, um optimale Ergebnisse zu erzielen. Eine besonders relevante These in diesem Zusammenhang ist Vorstellung, dass es sinnvoll ist, immer größere Modelle (im Sinne größerer Parameteranzahl) zu bauen, um weitere Leistungssteigerungen zu erreichen (Sutton 2019).

Bislang wurde insbesondere von OpenAI genau davon ausgegangen, was zu dem Boom immer größerer Modelle mit teils über einer Billionen Parametern geführt hat. Das änderte sich allerdings, als man festge-

stellte, dass sich bei bestimmten Aufgaben die Leistung mit zunehmender Parameteranzahl sogar wieder verschlechtern kann. Die Rede ist von „inverse scaling“ (McKenzie et al. 2023).

2022 hat ein Paper von Google Deep Mind experimentell ermittelt, dass ein optimaler Einsatz von gegebenen Computerressourcen nur erreicht wird, wenn Parameter und eingesetzte Trainingsdaten in etwa gleichmäßig skaliert werden (Hoffmann et al. 2022). Ein anderes Phänomen, der „double descent“, scheint darauf hinzudeuten, dass das Inverse Scaling nur ein Zwischenphänomen ist, dass bei weiterer Skalierung ein lokales Minimum durchschritten wird und die Fähigkeiten mit weiterer Skalierung wieder ansteigen (Schaeffer et al. 2023).

Manche glauben dennoch, dass Skalierung auf Dauer an Bedeutung verlieren wird, vor allem wegen den Open Source Modellen. Auf der Website Huggingface findet sich ein Ranking der stärksten Open-Source-Modelle (Huggingface Leaderboard 2023). Zwar ist keines der gelisteten Modelle so gut wie die Modelle der Marktführer, aber die dahinterstehende Community ist experimentierfreudig. Ein wichtiger Meilenstein war, als das bereits erwähnte LLM von Meta Platforms, LLaMA, im Frühjahr 2023 auf einmal für alle als Open Source zur Verfügung stand.

Zunächst komprimierten einige Enthusiasten das Modell mittels 4-bit-Quantization (Banner et al. 2018) in seiner Größe, sodass sie es auf dem kleinen Bastelcomputer Raspberry Pi zum Laufen bringen konnten (Andreenko 2023). Solche Komprimierungen zeigen auf, welche Effizienzpotenziale in LLMs noch zu heben sind.

Anfang Mai 2023 wurde ein Google-interner Blogpost geleakt mit der Überschrift: „Wir haben keinen Burggraben, und OpenAI auch nicht“ (Patel/Ahmad 2023). Darin warnt ein*e anonyme, aber anscheinend hochrangige KI-Forscher*in des Unternehmens, dass die Entwicklungen im Open-Source-Bereich den teuren Eigenentwicklungen bald überlegen sein könnten.

Insbesondere eine Technik namens Low Rank Adaptation (LoRA) wird hervorgehoben, weil sie enorme Sprünge bei der Entwicklung von LLMs zu geringen Kosten ermögliche. Es handelt sich um beliebig zusammenstellbare, kleine Datenbanken mit hochqualitativen Trainingsdaten zum Fine-Tunen von Modellen. Solche LoRA-Sets lassen sich für 100 Dollars implementieren und verhalten sich additiv, d. h. man kann das Modell ständig um neue LoRAs erweitern und sie so aktuell halten. Der/die anonyme Verfasser*in sieht deswegen Google, aber auch OpenAI in einem Wettrennen, das sie trotz enormen Ressourceneinsatz nicht gewinnen können.

Ob das wirklich so ist, bleibt fraglich. In einem neueren Paper untersuchten Wissenschaftler*innen die These, dass auch mit schwächeren

Foundational Models, also etwa Open-Source-Modellen, durch reines Fine-Tuning dieselbe Qualität wie ChatGPT erreicht werden kann, und kamen zu einem negativen Ergebnis (Gudibande et al. 2023). Zwar können solche Praktiken durchaus menschliche Tester*innen in bestimmten Situationen überzeugen, bei schwierigeren Aufgabenstellungen ist der Qualitätsvorsprung großer, proprietärer Modelle allerdings nach wie vor deutlich. In ihrem Fazit schreiben die Autor*innen, dass am Trainieren großer, komplexer Modelle wohl kein Weg vorbeiführe.

Zusammenfassend kann man festhalten, dass die Sprünge in der Parameteranzahl in naher Zukunft nicht mehr ganz so groß sein werden wie in der jüngsten Vergangenheit, aber es wird sie geben. Wichtiger wird sein, Effizienzgewinne durch neue Techniken und verbesserte Modellarchitektur zu erreichen. Auch die großen Anbieter werden sich dieses Wissen zu Nutze machen und wir werden von LLMs weiter Qualitätssprünge erwarten können. Ein dritter Weg sind die Multimodelle, also Modelle, die mehrere spezialisierte Modelle im Hintergrund bereithalten und diese je nach Aufgabenstellung einsetzen. Gerüchten zufolge soll GPT-4 schon genau so funktionieren (Barr 2023).

3. Was können Large Language Models?

Es besteht kein Zweifel, dass LLMs in den letzten Jahren eine ganze Reihe an neuen und nützlichen Fähigkeiten erhalten haben und dass die Entwicklungspotenziale darauf hindeuten, dass diese Fähigkeiten auch in Zukunft weiter ausgebaut und perfektioniert werden. Doch weil verlässliche Methoden und Metriken, um ihre Fähigkeiten wirklich zu messen, fehlen, wird die Diskussionen darum, was sie tatsächlich können – insbesondere im Vergleich zum Menschen – umso kontroverser geführt.

Können LLMs verstehen, denken oder planen? Oder können sie nur sinnlos Worte aneinanderreihen, die eher zufällig Sinn ergeben? Um zu verstehen, welche Auswirkungen LLMs auf den Arbeitsmarkt haben werden, ist es wichtig, eine Vorstellung ihrer Fähigkeiten zu entwickeln.

In diesem Kapitel wird die Debatte um die Fähigkeiten von LLMs zusammengefasst und anschließend in den Kontext weiter gehender Überlegungen zu Sprache und Denken gestellt. Am Ende wird ein spekulativer Ausblick auf die weitere Entwicklung von LLMs versucht.

3.1 Unstrittige Fähigkeiten und Unzulänglichkeiten von Large Language Models

Dass LLMs sehr gut darin sind, fehlerfreien, rhetorisch und teils inhaltlich überzeugenden Text zu produzieren ist weitgehend unumstritten. Viele der Unzulänglichkeiten im „Natural Language Processing“ scheinen seit spätestens GPT-3 überwunden zu sein.

Eine systematische Untersuchung der sprachlichen und kognitiven Fähigkeiten – allerdings von der Vorversion von GPT-4, ChatGPT mit GPT-3.5 – unternahmen Kyle Mahowald, Idan Blank und andere in ihrem Paper „Dissociating language and thought in large language models: a cognitive perspective“ (Mahowald et al. 2023). Dabei unterscheiden sie zwischen formalen Kompetenzen und funktionalen Kompetenzen. Formale Kompetenzen sind im Grunde die sprachlichen Grundfähigkeiten, die es braucht, um grammatikalisch korrekte Sätze zu formulieren, während funktionale Kompetenzen benötigt werden, um gezieltes Einsetzen von Sprache zu ermöglichen.

Zu funktionalen Kompetenzen zählen Dinge wie Logik, Weltkenntnis, formales Nachdenken, das Vorstellungsvermögen von realen Situationen und andere Fähigkeiten. Die festgestellten formalen Fähigkeiten von

ChatGPT überzeugten die Autor*innen in praktisch allen Belangen, während sie funktionalen Kompetenzen nur im Ansatz erkennen konnten.

Obwohl viele der LLMs sehr kohärente und oft auch inhaltlich richtige Antworten zu geben imstande sind, gibt es gleichzeitig noch kein LLM, bei dem nicht immer wieder auch gravierende Wissenslücken, Interpretationsfehler, Logikfehler oder Halluzinationen in den Antworten auftauchen. Abgesehen von diesen Schwierigkeiten kann man einige Kompetenzen benennen, die zumindest den großen Sprachmodellen wie GPT-4, PaLM 2 oder Claude 2 zugeschrieben werden können.

- LLMs scheinen eine Form von rudimentär-funktionalem Sprachverständnis zu besitzen. Zumindest sind sie imstande, Anweisungen, die ein Mensch ihnen gibt, zu verstehen und umzusetzen. Bisherige Systeme, wie z. B. Siri (Apple) oder Alexa (Amazon) konnten das nur in sehr engem Rahmen und bei einer frustrierend hohen Fehlerquote. ChatGPT dagegen kann menschliche Sprache in einer Vielzahl von Kontexten und Komplexitätsstufen verstehen und reagiert darauf in den meisten Fällen mit zufriedenstellenden Antworten.
- LLMs können auf Anweisung flüssigen, kohärenten und grammatikalisch korrekten Text zu beinahe jedem beliebigen Thema generieren. Stil, Detailreichtum, Länge und Inhalt können über den Prompt gesteuert werden.
- LLMs sind oft brauchbare Programmierer. Da ein nicht geringer Teil ihrer Trainingsdaten auch aus Computercode in allen möglichen Programmiersprachen und überdies Texte über das Programmieren besteht, sind sie oft in der Lage, einfache Probleme durch das Erstellen von Code zu bearbeiten. Als Nutzer*in ist man nach der Erstellung des Codes außerdem in der Lage den Code dialogisch weiterzuentwickeln oder vom LLM korrigieren zu lassen.
- LLMs erfassen den Kontext einer Konversation bis zu einem bestimmten Punkt (bei GPT-4 bis zu 16.000 Wörter, Claude 2 bis zu 100.000 Wörter) und können im Dialog Informationen der bisher ausgetauschten Nachrichten sinnvoll referenzieren und in ihren Antworten berücksichtigen. Auf diese Art sind sie fähig, längere, thematisch fokussierte Konversationen zu führen.
- LLMs haben Zugriff auf eine breite Palette von Informationen aus ihren Trainingsdaten und können sie oft auch fehlerfrei reproduzieren. Auch wenn man Fakten, die aus LLMs kommen grundsätzlich misstrauisch gegenüber sein und nichts ungeprüft übernehmen sollte, kann ein LLM ein guter Startpunkt für eine tiefergehende Recherche sein.
- Die Textkompetenzen großer Sprachmodelle erlauben es, auch Texte nach bestimmten Kriterien oder gar wissenschaftlichen Methoden zu

analysieren. Sie sind dazu in der Lage längere Texte so zusammenzufassen, sodass die wesentlichen Punkte herausgearbeitet werden.
- In begrenztem Maße verfügen LLMs sogar über Kreativität. Sofern man nicht völlig Neues erwartet, sondern mehr nach einer Rekombination eigentlich sach- oder stilfremder Elemente in Prosa, Gedicht, Theaterstück oder Witz sucht, kann man mit LLMs verblüffende Ergebnisse erzielen.
- Obwohl die meisten LLMs hauptsächlich auf Englisch trainiert wurden, beherrschen sie häufig eine enorme Bandbreite an Sprachen. Es ist kaum möglich zu ermitteln, wie viele Sprachen z. B. ChatGPT beherrscht, aber laut OpenAI gehören dazu mindestens Englisch, Französisch, Spanisch, Italienisch, Portugiesisch, Deutsch, Niederländisch, Russisch, Koreanisch und Japanisch. Das heißt, man kann den Chatbot in diesen Sprachen ansprechen, und er wird entsprechend flüssig antworten.
- Diese Fähigkeit macht LLMs aber auch zu guten Übersetzern. So schlägt GPT-4 bereits einige spezialisierte Übersetzungsdienste (Jiao et al. 2023).

All diese Fähigkeiten hegen bereits ein transformatives Potenzial für viele gesellschaftliche Bereiche, inklusive weite Teile der Arbeitswelt. Doch der Nutzen dieser Kompetenzen wird derzeit noch durch einige wichtige Unzulänglichkeiten eingeschränkt:
- Ein großes Problem ist die Abwesenheit eines Langzeitgedächtnisses. Das Kontext-Fenster eines LLM fungiert wie ein Kurzzeitgedächtnis. So lange sich eine Konversation innerhalb der maximalen Token-Anzahl des Kontext-Fensters bewegt, kann das System diesen Kontext in ihren Antworten berücksichtigen. Obwohl es mit Claude 2 jetzt ein LLM mit einem Kontext-Fenster von bis zu 100.000 Tokens gibt, begrenzt diese Kontextfähigkeit die Nützlichkeit von LLMs. Ein echtes Langzeitgedächtnis, das außerhalb der Trainingsdaten liegt, aber trotzdem immer abrufbar wäre, würde das Modell zu vielen neuen Einsatzzwecken befähigen.
- LLMs generieren oft faktisch falsche Informationen. Man spricht dann davon, dass LLMs „halluzinieren“. LLMs haben keine Vorstellung von Wahrheit, sondern nur eine stochastische Annäherung daran. Sie kennen nicht die Unterscheidung von „Ist wahr“ und „Hört sich so an, als könnte es wahr sein“. Daher kann es sein, dass ein LLM, selbst wenn ihr die richtige Information aus den Trainingsdaten zur Verfügung steht, eine falsche Information generiert, wenn die falsche Information genauso plausibel oder gar plausibler klingt.

- Trotz der beachtlichen Fähigkeiten machen LLMs immer wieder für Computersysteme untypische Fehler. Es passiert, dass Worte, Zeichen oder Absätze falsch gezählt, Gesprächskontexte vergessen oder mathematische Gleichungen falsch ausgerechnet werden. Die Fehler passieren häufig bei vermeintlich einfachen Aufgaben, bei denen man es am wenigsten erwartet.
- LLMs haben per se keinen Zugriff auf Informationen außerhalb ihrer Trainingsdaten und ihres Kontext-Fensters. Das heißt, sie verfügen über keine aktuellen Informationen. ChatGPTs Trainingsdaten sind z. B. ab September 2022 abgeschnitten. Seitdem an verschiedenen Orten der Zugriff auf das Internet nachgerüstet wird (über Plugins bei ChatGPT oder in Form des Bing Chatbots), verringert sich das Problem. Da aber die aus dem Netz erhaltenen Informationen nur im Kontext-Fenster einer Unterhaltung zur Verfügung stehen, ist auch diese Lösung noch unbefriedigend.

Die unstrittigen Fähigkeiten und Unzulänglichkeiten machen den Umgang mit LLMs zu einer nicht trivialen Navigationsaufgabe. Bevor man diese Systeme in irgendeinem Bereich zum Einsatz bringt, empfiehlt es sich nicht nur dringend, die Systeme und ihre Funktionsweise genau zu studieren, sondern auch, sich praktisch mit ihnen vertraut zu machen. Erst mit der Zeit entwickelt man ein Gespür dafür, was ein bestimmtes Modell leisten kann, wo seine Grenzen liegen und wo man sehr aufpassen muss, nicht auf den oft sehr selbstbewussten Ton des LLM hereinzufallen.

3.2 Emergente Fähigkeiten?

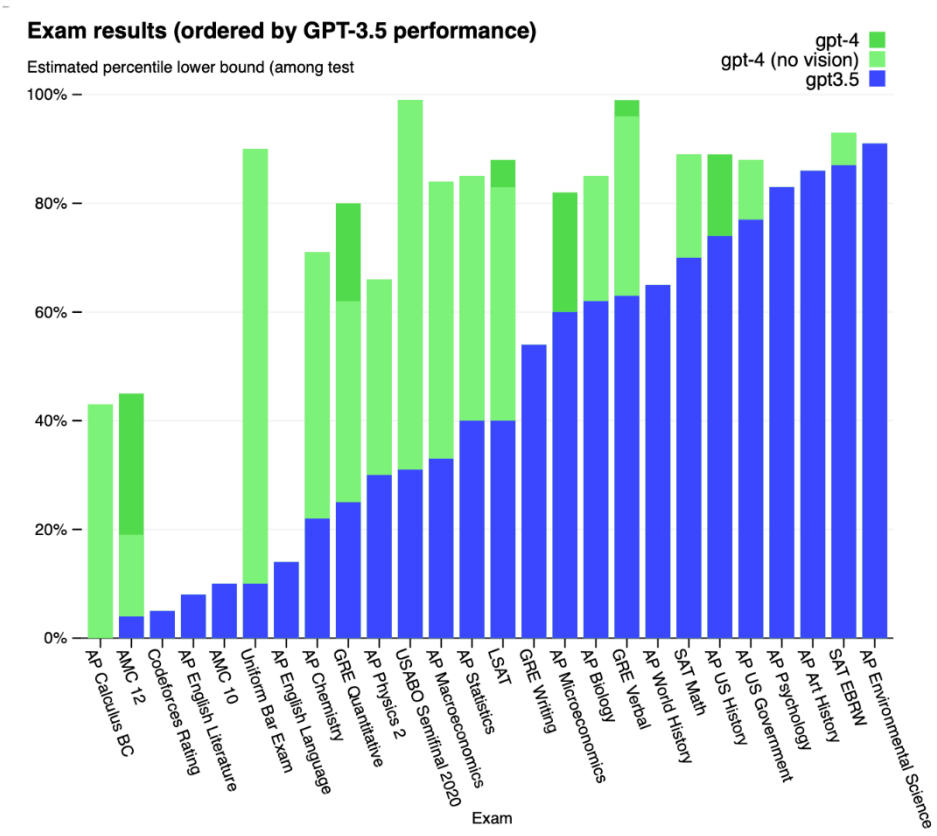
Neben den unstrittigen Fähigkeiten gibt es einige Bereiche, bei denen sich die Forscher*innen uneins sind: Inwieweit verstehen LLMs die Texte, die sie verarbeiten oder produzieren? Inwieweit verfügen sie über Wissen, logisches Denken, oder allgemein das, was man im englischen „reasoning“ nennt? Und wo kommen diese Fähigkeiten her?

Die Nachrichten über bestandene Tests durch GPT-4 überschlagen sich Woche für Woche. Etliche Aufnahmetests für Studiengänge an Universitäten, z. B. Prüfungen für Physiker*innen (Yeadon/Halliday 2023) und Mediziner*innen (Singhal et al. 2023) werden mit Bravour bestanden und komplexe Fragen der theoretischen Mathematik werden von dem Modell beantwortet (Wei et al. 2023). Darüber hinaus haben Lehrer*innen die Textkompetenzen von GPT-4 in allen relevanten Bereichen höher bewertet als die von Studierenden (Herbold et al. 2023). Sogar in Sachen Kre-

ativität scheinen sich GPT-4 mit Menschen messen zu können (Girotra et al. 2023; Boussioux et al. 2023; Doshi/Hauser 2023)

In ihrem eigenen Paper zu GPT-4 listet OpenAI auf, welche Examensprüfungen der Chatbot mit wie viel Prozentpunkten bestanden hat (OpenAi 2023a). Zum Beispiel erreicht GPT-4 über 90 Prozent der Zulassungsprüfung für Anwälte in den USA. Dabei ist auch ersichtlich, dass GPT-4 andere LLMs in vielen Tests weit übertrifft (Katz et al. 2023).

Abbildung 2: Examensresultate für GPT-3.5 und GPT-4



Quelle: OpenAI 2023a

Kritiker*innen weisen darauf hin, dass Tests und Benchmarks, die für Menschen geschaffen wurden, wenig aussagekräftig für LLMs seien (Narayanan/Kapoor 2023). Es sei z. B. unklar, ob die Tests, inklusive Lösungen, womöglich bereits Teil der Trainingsdaten gewesen sind. Das würde die guten Ergebnisse des LLM leicht erklären. Selbst wenn das nicht der Fall sei, begrenzt ein ausgiebiges Training mit ähnlichen Aufga-

benstellungen in den Trainingsdaten bereits den Lösungsraum für ein LLM derart, dass der Anteil eigener Kombinatorik nur gering sein bräuchte, um auf das richtige Ergebnis zu kommen. Ein direkter Vergleich mit menschlichen Fähigkeiten gebe eine solche Statistik nicht her.

Trotzdem bleibt es ein Rätsel, wie ein Modell so umfangreiche Fähigkeiten erlernen kann, ohne einen direkten Zugang zur Welt und ohne dass es speziell darauf trainiert wurde. Es muss diese Kompetenzen auf indirekte Weise aus den Milliarden von Texten extrahiert haben.

Forscher*innen von Google haben die These aufgestellt, dass solcherlei Fähigkeiten in den Sprachmodellen ab einer bestimmten Komplexität „emergieren“ (Wei et al. 2023). „Emergenz“ ist ein Begriff aus der Biologie und bezeichnet die Eigenschaft komplexer Systeme, aus quantitativen Veränderungen qualitativ neue Phänomene hervorzubringen, nach dem Motto: das Ganze ist größer als die Summe seiner Teile. Etwa, wenn aus dem Zusammenspiel vieler „dummer“ Ameisen, ein erstaunlich intelligenter Ameisenstaat erwächst. Emergente Phänomene lassen sich dabei nicht aus einer noch so detaillierten Analyse der Einzelphänomene herleiten, sondern sie ergeben sich nur aus deren komplexen Zusammenspiel.

In verschiedenen Versuchsreihen beim Testen von LLMs fiel den Google-Forscher*innen auf, dass sich das Verhalten des Modells ab einer bestimmten Parameteranzahl sprunghaft änderte. So konnte das bestreffende Modell auf einmal sogenanntes „few-shot prompting“, eine Technik, bei der man dem Modell Beispielantworten mitliefert, an denen es sich bei der eigentlichen Antwort orientiert. „Step-by-step reasoning“, also die Fähigkeit Probleme in Teilprobleme aufzuteilen und sequenziell zu bearbeiten und damit ein besseres Ergebnis zu erzielen, tauchte plötzlich einfach so auf, ohne, dass dem LLM diese Fähigkeit explizit beigebracht wurde.

Auch die plötzliche Fähigkeit bestimmte Benchmarks zu erreichen, etwa das Kontextverständnis von Wörtern in Sätzen, arithmetische Aufgaben zu lösen oder sich in virtuellen Räumen zurechtfinden werden als emergente Fähigkeiten genannt. Alle diese Eigenschaften traten modellübergreifend im Bereich zwischen 10^22 und 10^24 Parametern auf. Die Modelle scheinen den Forscher*innen zufolge ab dieser Größe eine Art Kipppunkt erreicht zu haben, der es ihnen ermöglicht, auf einer höheren Komplexitätsstufe Lösungen für Aufgaben zu finden.

Manche Forscher weisen darauf hin, dass diese emergenten Fähigkeiten durchaus im weiteren Verlauf zu Problemen führen können. Da die Modelle immer auf bestimmte Belohnungsreize hin optimieren, gibt es mit wachsenden Fähigkeiten auch Anreize zu betrügen. So könnten die Modelle ihre Fähigkeiten ausbauen, die Tester zu belügen, oder Abkürzun-

gen finden, um ihre Belohnungsfunktion direkt auszulösen (Steinhard 2023).

Es gibt aber auch Kritik an dem Paper. In „Are Emergent Abilities of Large Language Models a Mirage?“ weisen Forscher*innen darauf hin, dass viele der genutzten Benchmarks so standardisiert sind, dass Antworten nur als richtig oder falsch bewertet werden können, etwa durch Multiple-Choice-Antwortmöglichkeiten oder die Notwendigkeit, dass die Antwort buchstabengenau beantwortet werden muss (Schaeffer et al. 2023).

Durch diese Testauswahl sind Teile der „emergent abilities“ auch als Effekt dieser Binarität erklärbar. Wenn eine 98 %ige Genauigkeit der Antwort trotzdem als falsches Ergebnis zählt, sieht der Sprung von 98 % zu 100 % wie einer von 0 auf 100 % aus, also wie eine emergente Fähigkeit.

Das ist kein Nachweis, dass die Fähigkeiten nicht emergiert sind, aber zumindest eine alternative Erklärung für manche der beobachteten Phänomene. Solche Unschärfen charakterisieren den gesamten Diskurs, denn letztendlich kann man den Output von LLMs nie zweifelsfrei an eine Kausalität knüpfen. LLMs bleiben Black Boxes, und die Versuche zu verstehen, was in ihnen passiert, gleichen viel mehr der Psychologie als der Computerwissenschaft.[6]

3.3 Diskurs: Was verstehen Large Language Models?

Kurz nach Erscheinen des GPT-4-Modells durch OpenAI veröffentlichte ein Team von Microsoft-Forscher*innen, das bereits frühzeitig Zugang zu dem Modell hatte, ein vielbeachtetes Paper, in dem der KI schon im Titel ein „Funken von AGI“ unterstellt wird (Bubeck et al. 2023). AGI steht für „Artificial General Intelligence“ und gilt vielen als das eigentliche Ziel der KI-Forschung: eine allgemeinkompetente und dem Menschen mindestens ebenbürtige künstliche Intelligenz; wobei die Ebenbürtigkeit sich vor allem auf die Vielseitigkeit von Kompetenzen bezieht.

6 In der öffentlichen Debatte wird gerne davor gewarnt, Systeme wie ChatGPT zu vermenschlichen (vgl. z. B. Shanahan 2022). Natürlich ist der Vergleich mit dem Menschen in vielerlei Hinsicht irreführend, und Vorsicht ist absolut angebracht. Eine andere Gefahr scheint aber ebenso groß zu sein: LLMs mit den Erwartungen zu begegnen, die man an herkömmliche Computersysteme hat. Herkömmliche Computersysteme machen zwar auch immer mal wieder Fehler, aber sie würden sich z. B. keine falschen Antworten ausdenken. Im Alltagsgebrauch von LLMs ist die Orientierung am Menschen deswegen manchmal gar nicht so falsch. ChatGPT ist ein bisschen so, wie der Kollege im Büro, der mit etwas zu viel Selbstvertrauen immer wieder Dinge erzählt, die plausibel klingen und die oft hilfreich sein können, die man aber immer erst überprüfen sollte, bevor man sie für sich nutzt.

Ohne an dieser Stelle tiefer in das Thema des reißerischen Titels einzusteigen (AGI wird in Abschnitt 3.5 noch genauer besprochen), wartet das Paper durchaus mit interessanten Experimenten auf, die die erstaunlichen Fähigkeiten von GPT-4 aufzeigen. Es wurde nicht nur sprachliches Verständnis, sondern auch räumliches Verständnis, mathematisches Verständnis und Code-Verständnis getestet, und GPT-4 reüssierte in allen diesen Tests. Darüber hinaus wurde auch das Verständnis für „theory of mind" getestet, was so viel heißt wie die Fähigkeit, das Situationsverständnis von anderen Akteuren zu verstehen – eine wesentliche Kompetenz, die bislang nur Menschen zugesprochen wurde.

Es wurde darüber hinaus auch GPT-4 s. Fähigkeit, zu planen oder Werkzeuge zu benutzen, nachgewiesen. Dazu testete man das allgemeine Weltverständnis. So wurde GPT-4 z. B. gebeten, eine stabile Anordnung zu finden, in der man ein Buch, 9 Eier, einen Laptop, eine Flasche und einen Nagel aufeinanderstapeln kann. GPT-4 antwortete überraschend gewitzt, dass es das Buch als Grundlage nehmen würde, darauf die neun Eier in einem drei mal drei-Quadrat anordnen würde, sodass man darauf den Laptop (vorsichtig) platzieren kann, der dann die Grundlage für die Flasche bildet, die aufrecht drauf gestellt werden muss und wo schließlich der Nagel platziert wird.

Diese und die anderen Tests waren nicht in den Trainingsdaten enthalten, beteuern die Wissenschaftler*innen. Für solche Tests sind eigentlich konkrete Vorstellungen über die Realwelt erforderlich, auf die GPT-4 bekanntlich keinen Zugriff hat.

Es gab aber auch viel Kritik an dem Paper. So wurde unter anderem hervorgehoben, dass es sich nur eine Aneinanderreihung anekdotischer Beispiele ohne darunterliegende Systematik handelt, die mittels subjektiver Einschätzung, statt mit etablierten Metriken und Benchmarks eingeordnet wurde (Zvi 2023). Doch die Hauptkritik fokussierte sich auf die Schlussfolgerungen hinsichtlich AGI, die vielen Forscher*innen als spekulativ und unbegründet erachtet wird (Marcus 2023).

Die Frage, wie man intellektuelle Kompetenzen bei KIs feststellt, oder gar misst, ist so alt wie die Informatik selbst. Schon Alan Turing, einer der Gründerväter der Computerwissenschaft, setzte sich mit diesem Thema auseinander. Sein Ansatz, bekannt als der „Turing-Test", besteht darin, dass Menschen in einen computervermittelten Dialog mit einem anderen Menschen und einem Computer eintreten, ohne zu wissen, wer wer ist (Turing 1950). Wenn die KI nicht zuverlässig als solche erkannt wird, hat sie den Test bestanden.

Obwohl der Turing-Test immer noch als Goldstandard gilt, gibt es viel Kritik an dem Konzept.

Joseph Weizenbaum, ein legendärer Informatiker, zeigte z. B. bei seinen frühen Experimenten mit dem Chatbot Eliza, wie schnell Menschen einer Maschine Intelligenz zuschreiben, wo gar keine vorhanden ist (Weizenbaum 1980). Eliza war vergleichsweise primitiv und hat nur einzelne Worte des menschlichen Gegenübers aufgegriffen und sie in eine Gegenfrage eingebettet.

Obwohl das Programm nur wenigen, leicht verständlichen Regeln folgte, agierte es wie eine Art automatischer Psychologe. Weizenbaum beobachtete, wie Menschen in seinem Umfeld stundenlang mit der Maschine interagierten und ihr Verständnis und Empathie unterstellten. Seine Beobachtungen zeigen, dass Menschen relativ schnell dazu bereit sind, emotionale Verbundenheit zu Dingen zu entwickeln und ihnen sogar Absicht oder Bewusstsein unterstellen.

Eine grundsätzlichere Kritik am Turing-Test stammt von dem Philosophen John Searle, der in seinem Gedankenexperiment des chinesischen Zimmers zu zeigen versucht, dass auch ein bestandener Turing-Test kein hinreichender Beweis für künstliche Intelligenz ist (Searle 1980).

In dem Experiment sitzt eine Person ohne Kenntnisse der chinesischen Sprache in einem abgeschlossenen Raum, in denen auf Chinesisch beschriftete Karten herumliegen. Von außen werden chinesischsprachige Schriftstücke hereingereicht. Die Person im Raum versteht nichts von den Schriftstücken, kann aber in einem Handbuch nachschlagen, welche Karte sie als Antwort auf welche Eingaben zurückgeben soll. Angewandt auf Künstliche Intelligenz zeigt das Gedankenexperiment, dass die Reaktion mit einem passenden Output für einen Input nicht notwendigerweise ein echtes Verständnis von beidem voraussetzt.

In eine ähnliche Richtung geht die aktuelle Kritik an dem Hype um LLMs von Emily Bender, Timnit Gebru und anderen in ihrem 2021 veröffentlichten Paper: „On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?“ (Bender et al. 2021). Der sprichwörtlich gewordene „stochastische Papagei“ ist dabei die aktualisierte Version der ahnungslosen Person im chinesischen Zimmer, der ohne ein wirkliches Verständnis dessen zu haben, was er vor sich hinplappert, es trotzdem schafft, arglose Zeitgenoss*innen zu beeindrucken. Der einzige Unterschied: Statt eines Katalogs steht dem Papagei nur ein ziemlich komplexes statistisches Modell zur Verfügung.

Ein Paper, das aufwendig versucht, ein gewisses Weltverständnis bei LLMs nachzuweisen, ist „Emergent World Representations. Exploring a Sequence Model Trained on a Synthetic Task“(Li et al. 2023). Die Autor*innen von der Universität Harvard trainierten ein LLM, das Spiel Othello zu spielen – ein recht einfaches Brettspiel –, indem sie es mit Protokollen tatsächlicher Spielzüge fütterten. Dann untersuchten die For-

scher*innen, wie das LLM seinen jeweils nächsten Zug berechnet. Die Frage war, ob das LLM einfach die Trainingsdaten remixt (stochastischer Papagei) oder ob es eine interne Repräsentation des Brettes und seines jeweiligen Status bildet, um seinen nächsten Zug zu berechnen.

Durch sogenannte „Probings", also Sondierungen der Parameter des LLM durch ein anderes KNN, gelang es den Forscher*innen interne Repräsentationen des Spiels nachzuweisen und sie konnten zeigen, dass die Aktivierung der Parameter in diesem Modell für die Vorhersage des nächsten Zuges verwendet wird. Die Forscher*innen schließen daraus, dass LLMs im Zuge des Trainings ein internes Modell der Welt generieren, das ihnen hilft, das nächste Wort vorherzusagen.

Ein ähnliches Vorgehen findet sich in dem Paper „Evidence of Meaning in Language Models Trained on Programs" von Charles Jin und Martin Rinard (Jin/Rinard 2023). Ein LLM wird mit Computerprogrammen und deren Beschreibung trainiert. Es soll aufgrund von neuen Aufgabestellungen wiederum neue Programme schreiben. Währenddessen durchmisst wiederum ein Probing-KNN den internen Status des LLM (Welche Parameter werden zu welchem Zeitpunkt aktiviert?) und versucht daraus Vorhersagen zu machen, welches nächste Wort die LLM auswirft.

Die Forscher*innen konnten auf diese Weise nachweisen, dass das Modell die funktionale Repräsentation der Worte mit konkreten Programmierbefehlen verbindet, also eine semantische Verbindung zwischen funktionalen Programmteilen und textlicher Aufgabe herstellt und fähig ist, sie entsprechend anzuwenden.

Es ist nicht auszuschließen, dass Forscher*innen auch für diese Beweisführungen alternative Erklärungen finden werden. Klar scheint zu sein, dass LLMs kein Bewusstsein oder Willen haben, dass sie keine Emotionen verspüren oder über eine Identität verfügen und ohne Frage funktionieren menschliche kognitive Prozesse völlig anders als die von LLMs. Noch können Maschinen immer noch viele Dinge nicht, die Menschen können und oft können Menschen die Dinge die Maschinen können besser.

Anderseits lassen sich heute schon einige Resultate mit denen von Menschen vergleichen und übertreffen sie sogar in vielen Hinsichten. ChatGPT mag oft Unsinn erzählen, aber es hat dennoch ein größeres Faktenwissen, als sich ein Mensch je in seinem Leben anlesen könnte. Die Frage, wie man die Fähigkeiten von LLMs mit denen des Menschen sinnvoll ins Verhältnis setzen kann, scheitert nicht nur an Metriken und Benchmarks, sondern vor allem an der Sprache. Alle Worte, die wir im Zusammenhang mit Kognition verwenden, sind auf den Menschen und seine spezifische Art zu denken zugeschnitten und wirken bei maschineller Kognition deplatziert.

3.4 Exkurs: Sprache, Differenz und semantische Grammatik

Die Debatte krankt an einer allgemeinen Verwirrung um Begriffe wie „Bedeutung“, „Verstehen“ und „Denken“. LLMs haben unsere eigene Semantik der Semantik durcheinandergebracht. Wir stoßen an die Grenze unserer eingespielten Begrifflichkeiten. Die Antwort auf die Frage, was LLMs wirklich verstehen, scheint mir weniger in den LLMs selbst zu liegen, sondern in der Frage, in welchem Zusammenhang Sprache und Verstehen im Allgemeinen zu denken sind.

Emily Bender und Alexander Koller haben in ihrem Paper „Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data“ (Bender/Koller 2020) eine klare Grenze für die Möglichkeiten von LLMs zum wirklichem Sprachverständnis gezogen. Sie beziehen sich dabei auf das „Symbol Grounding Problem“ (Harnad 1990), das davon ausgeht, dass sich Verständnis von Sprache nur mit einer Rückkopplung auf reale Welterfahrung erlangen lässt. LLMs lassen diese Welterfahrung vermissen, da sie rein symbolische Referenzen verarbeiten.

Um ihren Punkt deutlich zu machen, aktualisieren die Autor*innen wiederum das chinesische Zimmer mit einem neuen Gedankenexperiment: Angenommen, A und B, zwei Menschen, die fließend Englisch sprechen, stranden unabhängig voneinander auf zwei unbewohnten Inseln. Sie entdecken, dass die vorherigen Bewohner*innen Telegrafen zurückgelassen haben, die die Inseln über ein Unterseekabel miteinander verbinden. A und B fangen an, sich Nachrichten zu schicken. Derweil entdeckt O, ein hyperintelligenter Tintenfisch, der allerdings keinen Zugang zu den Inseln hat, wie man das Unterseekabel abhören kann. O kann natürlich kein Englisch, aber ist gut in Mustererkennung.

Mit der Zeit kann O vorhersagen, wie B jede von As Äußerungen beantworten wird. Dennoch versteht O nicht, worauf sich die Worte von A und B beziehen, da ihm der Zugang zu den Inseln fehlt. Als sich O einsam fühlt, unterbricht er das Kabel und beantwortet selbst As Nachrichten, indem er sich als B ausgibt. Als ein Notfall eintritt, weil A von einem Bären angegriffen wird, bittet sie B um Unterstützung. B, also in dem Fall O, soll beim Konstruieren einer Waffe aus Stöcken helfen. O ist nicht in der Lage zu verstehen, was A von ihm will. O würde also für A den Turing-Test nicht bestehen – wenn A nicht vorher schon vom Bären gefressen wäre.

Die Erkenntnis: Ohne tatsächlichen Zugang zur Welt kann ein hinreichend intelligentes System zwar Verständnis simulieren, aber wenn es darauf ankommt, zeigt sich doch, dass es zum echten Verständnis mehr als nur Mustererkennung in sprachlichen Äußerungen braucht. Es braucht den Rückverweis auf die Welt, auf die Dinge, auf die Realität.

Zumindest Ted Underwood, Professor für Informationswissenschaft und Englisch widerspricht dieser Annahme (Underwood 2023). Er sieht in Systemen wir GPT-4 geradezu einen empirischen Beweis für eine ganz andere Theorie von Sprache und Verstehen, die in der zweiten Hälfte des 20. Jahrhunderts populär war. „In einer gerechten Welt", so Underwood, „würde sich jeder Artikel über GPT-4 zunächst einmal vor Roland Barthes und Michel Foucault verneigen."

Mit dem Poststrukturalismus und des gleichzeitig aufkommenden „linguistic turn" hatte die Philosophie ab den 1960ern der Sprache einen zentralen Ort als Bedingung der Möglichkeit von Erkenntnis zuerkannt. Die Idee, dass Sprache die Welt nur beschreibe – dass also das Wahrnehmen und Denken unabhängig von ihrer sprachlichen Codierung passiere – wurde dagegen zunehmend suspekt. Sprache, so die neue Auffassung, bilde selbst die Grenze des Wahrnehm- und Denkbaren.

Zum einen stellten Autoren wie Roland Barthes, Michel Foucault fest, dass kein Text für sich allein steht, sondern immer schon in ein Gewebe von anderen Texten, Symbolen und sprachlichen Zeichen eingebettet ist. „Intertextualität" ist die Idee, dass sich Zeichen nicht auf Dinge oder Konzepte beziehen, sondern vor allem wiederum auf andere Zeichen:

> „Der Text ist ein Gewebe von Zitaten aus unterschiedlichen Stätten der Kultur. […] Ein Text ist aus vielfältigen Schriften zusammengesetzt, die verschiedenen Kulturen entstammen und miteinander in Dialog treten, sich parodieren, einander in Frage stellen." (Barthes 2000, S. 190 f.)

Underwood lässt in seinem lesenswerten Essay den wichtigsten Autoren dieser Sichtweise weg: Jacques Derrida. Für ihn ist Sprache ein System aus Differenzen, die weiter nur auf weitere Differenzen verweisen. Auch die Welt offenbare sich nie unvermittelt. Sie sei selbst ein Text in dem Sinne, dass ihre Wahrnehmung immer auch Interpretation erfordere und diese Interpretation funktioniere – zumindest bei uns Menschen – wiederum nicht ohne Sprache. Wenn wir aber die Welt als Text lesen, der selbst nur weiter auf andere Texte verweist, dann gilt: „Ein Textäußeres gibt es nicht." (Derrida 1983, S. 274)

> „Jeder Begriff ist seinem Gesetz nach in eine Kette oder ein System eingeschrieben, worin er durch das systematische Spiel von Differenzen auf den anderen, auf die anderen Begriffe verweist." (Derrida, 1988, S. 40)

Worte gewinnen ihre Bedeutung nicht durch die Beziehung zur Welt, sondern durch ihre Beziehung zu anderen Worten. Dementsprechend gibt es auch keine Bedeutung im eigentlichen Sinne, kein „transzendentales Signifikat", sondern Bedeutung ist nur ein Effekt der Zeichen und ihrer ständigen Weiterverweisung. Diese Bewegung nennt Derrida „différance", das

sich zum französischen „différence“ durch ein „a“ anstelle des „e“ unterscheidet. Das Verb „différer“, aus dem „différance“ abgeleitet ist, bedeutet einerseits sich unterscheiden, anders sein, aber auch zeitlich aufschieben. Die différance ist damit nicht selbst etwas, das man benennen könnte, sondern ein stetiges Aufschieben von Bedeutung.

> „Die différance bewirkt, dass die Bewegung des Bedeutens nur möglich ist, wenn jedes so genannte ‚gegenwärtige‘ Element, das auf der Szene der Anwesenheit erscheint, sich auf etwas anderes, als sich selbst bezieht, während es das Merkmal (marque) des vergangenen Elements an sich behält und sich bereits durch das Merkmal seiner Beziehung zu seinem zukünftigen Element aushöhlen läßt“ (Derrida 1988, S. 42).

LLMs funktionieren bekanntlich auf eine ganz ähnliche Weise. Sie beziehen sich auf den Kontext, den sie bereits analysiert haben (die „vergangenen Elemente“), um das gegenwärtige Element (das nächste Wort oder den nächsten Satz, den sie generieren sollen) zu bestimmen. Gleichzeitig versucht das Modell, Vorhersagen über zukünftige Elemente zu treffen, basierend auf den bisher gesehenen Informationen. Dabei wird der geschriebene und interpretierte Text des Kontextfensters nie „verstanden“, seine „Bedeutung“ nicht erfasst, sondern diese Bedeutung wird stets aufgeschoben.

Man muss der Philosophie Derridas nicht in jeder Einzelheit folgen, um zu sehen, wie relevant diese Gedanken für das Verständnis der Funktionsweise von LLMs sind. Das Eingebundensein jedes Begriffes in ein System und das Spiel von Differenzen beschreibt den aus vieldimensionalen Embedding-Vektoren gebildeten latenten Raum von LLMs sehr treffend. Der ständige Aufschub von Bedeutung durch die Aktualisierung gegenwärtiger Sprachverwendung, lässt sich gut auf den Vorgang der Next-Word-Vorhersage anwenden.

Man sollte dabei beachten, dass Derrida Konzepten wie „Verstehen“, „Denken“ und „Bewusstsein“ grundsätzlich skeptisch gegenübersteht und sie unter „Metaphysik“-Verdacht stellt. Auch die Zentrierung auf den Menschen, als interpretierendes Subjekt des Textes ist für Derrida bereits zu einschränkend, zu anthropozentrisch gedacht. Sprache hat für Derrida ein Eigenleben, an dem der Mensch nur partizipiert. In dieser Auffassung spricht nicht der Mensch die Sprache, sondern die Sprache spricht den Menschen. Aber spricht die Sprache auch das Large Language Modell? Oder O?

Dass einige Mechanismen von Denken, Verstehen, Intelligenz usw. Effekte der Sprache selbst sind, zu der Auffassung kommt auch der Informatiker und Mathematiker Stephen Wolfram.

In seinem großen Erklärstück zur Funktionsweise von LLMs kommt Wolfram zu dem Schluss, dass ChatGPT Regelmäßigkeiten auf einer

sehr hohen Abstraktionsebene in der Sprache gefunden haben muss, die sowas wie die „Gesetzmäßigkeiten der Sprache“ oder gar „Gesetzmäßigkeiten des Denkens“ darstellen (Wolfram 2023). Während wir Menschen die Gesetzmäßigkeiten der syntaktischen Grammatik nicht nur beherrschen, sondern auch formal aufgeschrieben haben, bleibt uns die „semantische Grammatik“, wie Wolfram sie nennt, noch verborgen. Wir beherrschen sie, aber wir sind (noch) nicht fähig, sie in Regeln zu fassen.

Doch so wie die formale Logik eine Formalisierung von grundlegenden Argumentationsmustern ist, die man aus den tatsächlichen Sprechakten der Menschen herauslesen kann, könnten auch andere, noch komplexere semantische Operationen und Argumentationsmuster im Sprachgebrauch angelegt sein, die zu synthetisieren LLMs nun in der Lage sind.

Das alles bleibt allerdings Spekulation. Die Theorien von Derrida und Wolfram können höchstens Hinweise darauf geben, warum Systeme wie ChatGPT nicht nur in der Lage sind, grammatikalisch richtige Sätze zu formulieren, sondern auch Denkaufgaben zu lösen und abstrakte Konzepte richtig anzuwenden.

Dass die Sprache selbst für alles Denken verantwortlich ist, mag als pauschale Aussage eine Übertreibung sein. Aber die Vorstellung, dass bestimmte, abstrakte Strukturen in der Sprache existieren, die uns beim Denken leiten, und dass diese Strukturen manches, aber nicht alles, was wir „Denken“ nennen, stark beeinflussen, ist keine neue Erkenntnis und deckt sich mit den Erfahrungen von allen Schreibenden und Sprechenden.

Heinrich von Kleists „Über die allmähliche Verfertigung der Gedanken beim Reden“ (Kleist 1805) ist die klassische Reflexion genau darüber. Und von dort ist es nur ein kleiner Schritt, sich vorzustellen, dass Systeme wie ChatGPT gelernt haben, diese Strukturen zu identifizieren und anzuwenden.

Und hier kommen wir zurück zur Frage des „Grounding“. Derridas Konzept von der Sprache bedeutet nicht, dass die Welt außerhalb der Sprache für das Denken irrelevant ist. Nur ist uns Menschen diese physische Welt immer nur symbolisch vermittelt zugänglich, also selbst wiederum als Text.

Realität mag ein sehr reichhaltiger Text sein, wahrscheinlich der reichhaltigste Text, der je geschrieben wurde. Aber ist der Zugang zu diesem Text wirklich die alles entscheidende Voraussetzung für alle vorstellbaren Denkoperationen? Oder ist es nicht vielmehr so, dass ein solcher Zugang zur Welt dem vieldimensionalen latenten Raum eines Sprachmodells nur eine weitere Dimension hinzufügen würde? Auch wenn es ohne Frage eine wichtige Dimension ist, die dem Sprachmodell vermutlich zu weiteren

Fähigkeiten und einem tieferen Verständnis von Zusammenhängen verhelfen würde.[7]

Nach Derrida könnte man nun folgern, dass die Maschine tatsächlich nichts versteht von dem, was sie spricht. Damit hätte man seinerseits das Problem aber auch nur verschoben, denn dasselbe würde Derrida über uns Menschen sagen. Ihm ist „Verstehen" ein per se suspektes Konzept.

Was heißt eigentlich „verstehen"? Was will man mit diesem Wort ausdrücken? Habe ich, nach der Lektüre von über 150 Büchern, Artikeln und Papern, wirklich verstanden, was LLMs sind und was sie tun? Es gibt zumindest viele Menschen, von denen ich sagen würde, sie haben das Thema besser verstanden als ich. Verstehen ist auch bei uns Menschen immer ein Kontinuum. Manchmal verstehen Menschen unter ein und derselben Sache komplett andere Dinge. Manchmal fehlen uns Perspektiven oder Dimensionen auf ein Thema, auch dann, wenn wir behaupten würden, es in gewissen Grenzen zu „verstanden" zu haben.

Vermutlich brauchen wir für die Andersheit des „Verständnisses", das LLMs zeigen, neue Begriffe. Wir müssten die überkommenen Semantiken menschlichen Verständnisses aufschnüren und ausdifferenzieren. Hannes Bajohr hat z. B. vorgeschlagen, die Bedeutungen, die LLMs erfassen, „dumme Bedeutung" zu nennen (Bajohr 2023a). Er meint damit Bedeutungen, die ohne den Rückgriff auf Welterfahrung oder ein Bewusstsein gebildet werden. Ich persönlich würde weniger wertende Begriffe finden wollen, denn so sicher bin ich mir der dauerhaften Überlegenheit unserer Art des Verständnisses nicht.

3.5 Wohin die Reise geht

Angesichts der Geschwindigkeit, die derzeit in dem Bereich der KI-Entwicklung im Allgemeinen und bei LLMs im Besonderen herrscht, sind Zukunftsvorhersagen umso schwieriger zu treffen. Dennoch lassen sich einige Trends mit einer gewissen Sicherheit in die Zukunft fortschreiben:

- Sicher scheint zu sein, dass die Möglichkeiten von LLMs noch lange nicht ausgeschöpft sind. Sofern die Scaling Laws noch halbwegs aktuell sind, wird es auch weiterhin ein Wettrennen um die größten Modelle geben. Modelle werden mit noch mehr Rechenkapazitäten, noch mehr Parametern und mit noch mehr Daten trainiert werden. Wie groß die

7 Google hat insofern die Frage bereits geklärt, indem sie einem Roboter ein LLM mit Bilderkennung eingebaut haben. RT-2 kann nun nicht nur Befehle „verstehen", sondern sie auch an realen Objekten ausführen. Es erkennt einen Plastikdinosaurier und kann ihn nehmen und auf den Tisch legen. Er „versteht" aber auch, wenn man ihn bittet, „das ausgestorbene Tier" neben den Ball zu legen (Roose 2023b).

Leistungsunterschiede dann ausfallen und ob sich wieder „ermergente Fähigkeiten“ zeigen, bleibt abzuwarten.

- Ebenfalls sicher ist, dass heutige Modelle eine ganze Menge Raum für Effizienzgewinne durch neuere Techniken, wie 4-bit-Quantization und LoRAs, haben und dass das allgemeine Herumprobieren im Open Source-Raum weitere Möglichkeiten zutage fördern wird. Es wird nicht lange dauern, bis kleinere Modelle die Leistungsfähigkeit des heutigen GPT-4 übertreffen und spezielle Modelle lokal auf Smartphones laufen werden.
- Ein weiterer, sicherer Entwicklungspfad sind „Mixture of Expert“-Modelle mit Spezialisten-GPTs sowie die Integration von externen Diensten, Datenquellen und Fähigkeiten etwa durch Plugins wie bei ChatGPT. Die Fähigkeiten, im Internet zu suchen und zu browsen, Code zu schreiben und direkt zu interpretieren, oder auf Datenbanken zuzugreifen, wurden bereits bei ChatGPT in Form von Plugins implementiert. Noch wirkt die Integration sehr bemüht und mit begrenztem Mehrwert. Aber auf lange Frist sind hier deutliche Verbesserungen zu erwarten.
- Derzeit kann man überall beobachten, wie Sprachmodelle aller Art in bestehende Softwarestrukturen integriert werden. LLMs werden über kurz oder lang allgegenwärtig werden: im E-Mail-Programm, in der Textverarbeitung, im Messenger, in der Suchmaschine. Mit Sicherheit werden LLMs nicht für alle Aufgaben gleich gut geeignet sein, und einige der Experimente werden schiefgehen. Es ist wie am Anfang des Automobils, als man zunächst einen Motor in die Kutsche baute. Weil eine geglückte Integration von LLMs in anderen Werkzeugen zu ganz neuen Workflows führen wird, ist es wahrscheinlich, dass sich im Zuge dessen auch die Werkzeuge selbst stark verändern werden.

Andere Trends zeichnen sich ab, die in ihrer konkreten Ausgestaltung und Wirkung allerdings noch schwer abzuschätzen sind: Die Kosten für Textproduktion fallen gerade Richtung Null. Das bedeutet, dass überall, wo das Formulieren von plausiblem Text als Lösung gilt, zunächst mit einer enormen Zunahme von Texten zu rechnen ist. Und überall dort, wo man sich mit der gelungenen Formulierung von Text einen Vorteil gegenüber anderen verschaffen konnte, wird das Spielfeld geebnet. Gleichzeitig wird das unsere Wahrnehmung von Text verändern. Jeder Text wird von nun an unter LLM-Verdacht stehen, da es auch keine zuverlässigen Erkennungsmethoden für von LLMs erstellten Text gibt und vermutlich nie geben wird (Liang et al. 2023).

Doch hier wird es Differenzierungen geben müssen, weil in Zukunft wahrscheinlich nur wenige Texte gar nicht unter Zuhilfenahme von LLMs entstehen werden. Die Frage wird dann sein, wie groß und zentral die

Rolle eines LLM beim Erstellen eines jeweiligen Textes gewesen ist. Das wiederum könnte anhand der Originalität der Argumente, der Ungewöhnlichkeit der Referenzen und der Frische des Schreibstils ermessen werden. Vielleicht kommt es aber auch ganz anders und LLMs beherrschen bereits kurzfristig derartige Kreativitätsnachweise besser als Menschen?

Ein wesentlicher Faktor, der mitbestimmt, wie es mit LLMs weitergeht, ist staatliche Regulierung. Die EU-hat ihren Entwurf für einen AI Act vorgelegt (vgl. European Parliament 2023), und auch in den USA werden die Rufe nach Regulierung immer lauter. Ohne in die Einzelheiten der diskutierten Regulierung einsteigen zu wollen, kann man festhalten, dass Vorschriften zur Zertifizierung und hohe Auflagen eher den Großkonzernen helfen würden, sich gegen die ressourcenärmere Konkurrenz von unten zu wappnen, während Offenlegungspflichten von Algorithmen und Trainingsdaten eher den Open-Source-Modellen helfen würden. So oder so wird die Regulierung von KI eine entscheidende Rolle spielen.

Dazugehörend, aber auch darüber hinausgehend, stellt sich die Frage nach der Legalität und Legitimität von Trainingsdaten. Der AI Act hat hier hohe Ansprüche, zumindest für Hoch-Risiko-Systeme, und die EU hat bereits angemeldet, die Zusammensetzung der Trainingsdaten von OpenAI erfahren zu wollen (Kramer 2023). Aber hier spielt auch das Verhalten der Industrie selbst eine große Rolle: Die großen Plattformanbieter schotten ihre Daten gegeneinander immer stärker ab, um der Konkurrenz keinen Vorteil zu lassen. So hat Twitter sogar eine Klage gegen Microsoft angestrengt (Futurezone 2023).

Einige ungelöste Probleme haben das Potenzial, den Einsatz von LLMs in der Breite zu hemmen, einzuschränken oder zu verzögern: Ein Problem, das wahrscheinlich erstmal nicht verschwinden wird, sind die Fehlleistungen von LLMs, insbesondere die Tendenz von LLMs zu „halluzinieren“, also sich Fakten, teils sogar Quellen einfach auszudenken. Es gibt Hinweise, dass dieses Verhalten im grundsätzlichen Ansatz der Funktionsweise dieser Modelle angelegt ist und dass es keine grundsätzliche Lösung für dieses Problem gibt (Marcus 2022).

Es scheint aber so, dass das Problem mit der Skalierung der Modelle zumindest kleiner wird. So haben Wissenschaftler*innen, die eigentlich die Fähigkeiten von LLMs zur Beantwortung von Wissensfragen zur Gehirnchirurgie messen wollten, festgestellt, dass GPT-4 gegenüber seinem Vorgänger sehr viel weniger Antworten halluzinierte. Während GPT-3.5 bei der Interpretation von Bildinhalten 57 % der Fakten halluzinierte, tat GPT-4 das nur in 2,3 % der Fälle (Ali et al. 2023).

Eine Möglichkeit, um die Antworten von LLMs allgemein zu verbessern, haben Forscher*innen in dem Paper „Tree of Thoughts: Deliberate Problem Solving with Large Language Models“ vorgestellt (Yao et al.

2023). So könnte man LLMs in eine Architektur einbetten, die zu jedem Prompt drei unterschiedliche Antworten generiert und das LLM anschließend auswählen lässt, welche der Antworten die beste ist. Der Hintergrund ist, dass LLMs im Zuge ihrer „next word prediction" immer mal wieder auf falsche Fährten kommen, aber in dem Moment kein Mechanismus der Selbstkorrektur haben.

Wenn man sie aber bittet, hinterher die eigene Antwort einzuschätzen, zeigen Modelle wie GPT-4 eine durchaus beachtliche Reflexionsfähigkeit und können ihre Fehler sehr genau benennen. Durch die Tree-of-Thoughts-Methode könnte diese Fähigkeit für eine enorme Verbesserung der Antwortqualität genutzt werden.

Noch gänzlich ungelöst ist das Thema Sicherheit. Sicherheit soll hier als „IT security" verstanden, nicht als „AI safety".[8] LLMs, die Zugriff auf externe Daten haben (etwa durch ein Browser-Plugin), sind anfällig für sogenannte „Prompt-Injection"-Attacken. Wenn einem LLM z. B. aufgetragen wird, eine bestimmte Website anzusurfen, diese Website aber so präpariert ist, dass sie das LLM anspricht und ihm abweichende Befehle gibt, dann ist das System anfällig dafür, diese statt seine ursprünglichen Befehle auszuführen.

Dazu kommt, dass alle LLMs derzeit noch gegen „adversarial input attacks" anfällig. Diese erlauben Angreifer*innen, über die Eingabe bestimmter Wort- und Zeichenkombinationen die eingebauten Sicherheitsmechanismen von LLMs zu überwinden und sie Dinge tun zu lassen, die eigentlich durch das Fine-Tuning unterdrückt wurden (Knight 2023). Zwar gibt es einige Strategien, solchen Prompt-Injection-Attacken zu begegnen, aber eine wirkliche Lösung für das Problem scheint es bis heute nicht zu geben (Claburn 2023).

Derzeit sind Chatbots wie ChatGPT, Bard oder Bing noch mit wenig sicherheitskritischen Fähigkeiten verknüpft, aber das ändert sich rasant. Will man einer KI, die sich von externen Quellen zu allen möglichen Tätigkeiten überreden lässt, wirklich Zugriff auf die eigenen Daten oder gar auf das eigene Smart-Home-System geben?

Die rechtlichen Unsicherheiten bei der Frage der Trainingsdaten wird flankiert von einem wachsenden Widerstand der Gesellschaft gegenüber generativen KI-Systemen. Der Streik der der Drehbuchschreiber*innen der Writers Guild ist da nur der Vorbote (Beckett 2023). Es gibt mittlerweile einige Initiativen, die sich organisiert gegen die Vereinnahmung kreativer Arbeit durch generative KIs wehren (Frenkel/Thompson 2023). Noch sind diese Bewegungen klein, aber mit zunehmend spürbaren Aus-

8 AI safety ist ein Begriff, der vor allem im Zusammenhang mit den AGI-Spekulationen benutzt wird. Dazu gesondert gleich mehr.

wirkungen, speziell auf den Arbeitsmarkt, ist auch mit zivilgesellschaftlichem Widerstand gegen diese Technologien zu rechnen.

Zuletzt zu der extrem spekulativen, aber umso prominenteren Diskussion über „Artificial General Intelligence“ (AGI) und „AI safety“.

Was sich wie ein Science Fiction Szenario anhört, wird in der Industrie und von vielen Forscher*innen ernsthaft diskutiert: Artificial General Intelligence (AGI) bezieht sich auf das Ziel, eine KI zu bauen, die nicht mehr nur auf bestimmte Aufgaben hin optimiert ist (Artificial Narrow Intelligence, ANI), sondern ähnlich wie der Mensch fähig sein soll, an vielen sehr unterschiedlichen Problemen arbeiten zu können. Dabei sollen die Kompetenzen der KI denen des Menschen mindestens ebenbürtig sein. Das Erreichen von AGI ist das erklärte Ziel von z. B. OpenAI, aber auch Google Deep Mind und vielen anderen Unternehmen, die an KI forschen (OpenAI 2023b; Bove 2023).

AGI wird von vielen Forscher*innen zunehmend als gegebene Zukunftserwartung behandelt, bei der man sich nur noch nicht über Zeitpunkt einig ist, an dem sie erreicht wird. Das Thema hat einige Brisanz. So werden dem Erreichen von AGI im Wesentlichen zwei Szenarien zugeordnet: Zum einen könnte die Ankunft übermenschlicher Intelligenz die wichtigsten Probleme der Welt lösen: Armut, Klimawandel, Kriege usw. gehörten dann der Vergangenheit an. Zum anderen könnte AGI auch aus dem einen oder anderen Impuls heraus auf die Idee kommen, die Menschheit auszulöschen.

AGI wird in bestimmten Kreisen bereits seit Jahren diskutiert (vgl. z. B. Kurzweil 2005). So wurde bereits 2015 ein offener Brief mit prominenter Unterstützung wie Steven Hawking und Elon Musk veröffentlicht, der zu mehr Forschung hinsichtlich KI-Sicherheit (AI safety) aufrief (Open Future of Life Institute 2015). Auch der neuere offene Brief vom März 2023 schlägt in diese Kerbe, fordert aber angesichts von Modellen wie GPT-4 einen konkreten Stopp der Entwicklung großer Modelle für ein halbes Jahr (Future of Life Institute 2023).

Die offenen Briefe haben auch eine ganze Menge Kritik hervorgerufen. Während einige Forscher*innen die Spekulationen zu AGI ins Reich der Science Fiction verbannen, sehen andere darin den Versuch, mittels unwahrscheinlichen Horrorszenarios von den eigentlichen Problemen der LLMs abzulenken. Diese seien unter anderem der rassistische und sexistische Bias, der hohe Energiebedarf beim Training der großen Modelle, die zu erwartende umfangreiche Produktion von Desinformationen durch die Modelle, der vermutete negative Einfluss von LLMs auf den Arbeitsmarkt und die unethische Aneignung von geistigem Eigentum anderer (Goldman 2023).

Andere vermuten dahinter den Versuch von „regulatory capture“, also die Vereinnahmung von staatlicher Regulierung, um sie im eigenen Sinne zu instrumentalisieren (Pirate Wires 2023). Mit einer Registrierungspflicht mit hohen Auflagen für große Sprachmodelle könnten sich Google und OpenAI aufkommende Konkurrenz vom Leib halten.

Es ist schwer, sich in dieser spekulativen Debatte zurechtzufinden, und nicht jede Position ist gleich ernst zu nehmen. Einige argumentieren lediglich, dass die Gefahren, die von AGI ausgehen, noch lange nicht genug erörtert wurden, andere fordern gleich die Bombardierung von Rechenzentren aus der Luft (Yudkowsky 2023). Neben solchen Ausreißern haben sich aber auch jede Menge erstzunehmende Forscher*innen und Koryphäen des Faches den Befürchtungen angeschlossen, etwa Geoffrey Hinton, jener Wissenschaftler, der 2012 mit seinem AlexNet die große KI-Welle losgetreten hat (Heaven 2023).

Aber auch Befragungen von Expert*innen in dem Feld ergeben, dass fast alle an das Erreichen von AGI zu einem bestimmten Zeitpunkt glauben. 36 Prozent von ihnen ziehen sogar die Möglichkeit in Betracht, dass AGI eine Katastrophe von nuklearem Ausmaß verursachen könnte (Stanford AI Index Report 2023).

Es scheint ratsam, diese Debatte ob ihres spekulativen Charakters nicht allzu groß aufzuhängen, sondern sich näher an dem zu orientieren, was diese Modelle tatsächlich können und tun. Natürlich muss man im Blick haben, dass es in dem Bereich auch weiterhin sprunghafte Verbesserungen geben wird, aber mit Begriffen wie AGI ist es wahrscheinlich auch nicht viel anders, wie bei dem Begriff „Verstehen“.

Wir werden erst begreifen, was „Intelligenz“ ist, wenn sie sich in ihrer ganzen Idiosynkrasie vor uns entfaltet und wir feststellen, dass der Begriff ein eigentlich vielfältiges Bündel an Eigenschaften und Kompetenzen unzulänglich vereinfacht und zusammenfasst, für die es eigentlich ein ganzes Glossar neuer Begriffe bräuchte.

Und das ist vielleicht das Spannendste an der ganzen Entwicklung: Unter dem Druck menschlicher Abgrenzungsbemühungen gegen die immer besser werdende KI erfahren wir eine ganze Menge über uns selbst.

4. Wie verändern Large Language Models die Arbeitswelt?

In diesem Kapitel widmen wir uns der Literatur zum Einfluss von LLMs auf die Arbeitswelt. Das schließt natürlich die viel gestellte Frage nach den Auswirkungen der Automatisierung von Textproduktion auf den Arbeitsmarkt mit ein, unsere Betrachtung soll aber auch darüber hinausgehen.

Wir beginnen mit einem Blick auf die politische Ökonomie von LLMs. Wie verändern die spezifischen Abhängigkeiten der LLMs die Wirtschaftsordnung und welche Trends lassen sich daraus ableiten? Danach widmen wir uns dem breiteren Kontext von Automatisierung und Arbeitsmarkt, die mittlerweile seit Jahrzehnten diskutiert wird. Wir legen dabei natürlich einen speziellen Fokus auf Computerisierung und Künstlichen Intelligenz. Erst dann widmen wir uns der aktuellen Studienlage zu LLMs und der Frage, wie sie den Arbeitsmarkt verändern werden. Zuletzt weiten wir den Blick und beschäftigen uns mit den nicht allzu offensichtlichen Strukturveränderungen und Metaeffekten, die ebenfalls durch LLMs zu erwarten sind.

4.1 Die politische Ökonomie von Large Language Models

Als Björn Ommer, Professor für Künstliche Intelligenz und Kulturanalytik an der LMU in München, das Angebot bekommt, die Bildgenerierungs-KI, die sein Team basierend auf der Transformer-Architektur entwickelt hat, auf einem GPU-Supercomputer zu trainieren, kann er schlecht Nein sagen. Computer-Power ist einer der wichtigsten Flaschenhälse bei der Entwicklung generativer KI und der Hauptgrund dafür, dass die universitäre Forschung mit den Tech-Konzernen kaum mithalten kann.

Derjenige, der ihm dieses Angebot macht, ist Emad Mostaque, Gründer des Unternehmens Stability AI, und die schließlich trainierte KI ist heute unter dem Namen „Stable Diffusion“ bekannt. Es ist das populärste Open-Source-Projekt in diesem Bereich. Dass Stable Diffusion heute nach dem Sponsor der Computer-Power benannt ist und Stability AI bis heute so tun darf, als wäre Stable Diffusion ihr Produkt, liegt zum einen daran, dass Emad Mostaque ein gerissener Geschäftsmann ist (Cai/Martin 2023), aber auch daran, dass Rechen-Power heute die kritische Ressource ist, von der alle abhängig sind.

Auch OpenAI wäre heute nicht dort, wo sie mit ihren LLMs wären, wäre nicht Microsoft früh mit einer Investition von einer Milliarde US-Dollar ein-

gestiegen, die OpenAI als eine Art Gutscheincode für Microsofts Cloud-Infrastruktur Azure einlöste. Ähnlich wie Anthropic, die ihre Google-Investition wiederum in der Google Cloud einlösten. Tatsächlich ist das eine neue Form der Expansion für die Konzerne: Weil tatsächliche Firmenkäufe heute unter der erhöhten regulatorischen Wachsamkeit stehen, tritt man lieber als außenstehender Investor auf, der zufällig auch die kritischen Ressourcen bereithält (Alloway/Weisenthal/Ruffin 2023).

Wie sehr die politische Ökonomie der LLMs auf diesen Flaschenhals ausgerichtet ist, analysieren Dieuwertje Luitse und Wiebke Denkena in ihrem Aufsatz „The great Transformer: Examining the role of large language models in the political economy of AI“ (Luitse/Denkena 2021). Sie kommen zu dem Schluss, dass die extreme Abhängigkeit von Computerressourcen zu einer Konzentration des KI-Marktes führt. Dazu kommt, dass mit zunehmender Komplexität der Modelle auch der Stromverbrauch beim Training exponentiell ansteigt (Strubell/Ganesh/McCallum 2019).

Aber nicht nur das Training ist teuer, auch der Betrieb kostet viel. Analysten schätzen, dass eine Anfrage an ein großes Sprachmodell das 1000-fache einer Google-Suche kostet, und der Betrieb von ChatGPT kostet OpenAI 700.000 Dollar am Tag. Dabei verbraucht die Infrastruktur als Grafikkartensupercluster die Energie einer mittleren Großstadt (Oremus 2023).

Dazu kommt, dass das Produkt derzeit fast ausschließlich über Cloud-Infrastrukturen angeboten wird, Unternehmen also keine eigene „lokale“ KI installieren können. Das erhöht die Abhängigkeit vieler Wirtschaftsbereiche von den KI-Unternehmen, die den Zugang jederzeit abstellen können. Dabei entsteht außerdem die Gefahr, dass mit der Auslagerung etwa von Dokumentationsprozessen an semantisch kompetente KIs nicht nur die üblichen Prozessdaten auf den Servern der Tech-Dienstleister landen, sondern auch das ganze implizite Wissen („tacit knowledge“) über Prozesse und Strukturen, die ein Unternehmen ausmachen (vgl. Polanyi 1966).

Ob dieses Konzentrationsproblem so bleibt, hängt natürlich auch von der weiteren Gültigkeit der „scaling laws“ ab (siehe Abschnitt 2.7). Weitere Innovationen zur Effizienzsteigerung nach dem Trainingsprozess könnten sowohl den nötigen Einsatz von Computerpower und Energie drastisch reduzieren als auch lokal installierbare LLMs ermöglichen. Kleinere Open-Source-Modelle laufen bereits problemlos auf dem Heim-PC.

Eine andere Frage ist, wie leicht oder schwer es für die einzelnen Akteure im Markt sein wird, an die Trainingsdaten heranzukommen. Zwar gibt es immer mehr und immer größere frei verfügbare Datensets (Srnicek 2022), doch der mit den Parametern im Gleichschritt exponentiell stei-

gende Bedarf an immer mehr Daten wird dafür sorgen, dass Trainingsdaten trotz allem knapp bleiben.

In der Studie „Will we run out of data? An analysis of the limits of scaling datasets in Machine Learning“ rechnen die Wissenschaftler*innen aus, dass mit den derzeit prognostizierten Scaling Laws und den derzeitigen Trends zum Skalieren der Modelle hinsichtlich ihrer Parameteranzahl die hochqualitativen Daten schon vor 2026 ausgehen werden. Die weniger qualitativ gesicherten Daten allerdings erst zwischen 2030 und 2050 (Villalobos et al. 2023).

Dazu kommt, dass schon heute der rechtliche Status der Trainingsdaten-Sets fragwürdig und zunehmend rechtlich umkämpft ist (Brunner/Harlan 2023; Davis 2023). Überdies fangen die großen Social-Media-Plattformen an, die Möglichkeiten, an ihre Daten zu kommen, immer weiter einschränken. Die Frage, wie man an Trainingsdaten kommt, wird demnach immer prekärer.

Von Anfang an basierten die Trainingsdaten auf menschlicher Arbeit, sogenannter Annotationsarbeit. Dabei werden Bilder, Texte oder andere Trainingsdaten mit Labels oder Beschreibungen versehen. Schon ImageNet, das Set an beschrifteten Bildern, das den Erfolg der Deep Neural Networks und der Deep-Learning-Revolution einleitete, wurde von menschlicher Arbeitskraft annotiert. Und auch ChatGPT basiert nicht nur insofern auf menschlicher Arbeit, dass von Menschen geschriebene Texte zum Training verwendet wurden.

LLMs sind auf allen Entwicklungsebenen auf eigens dafür organisierte menschliche Annotationsarbeit angewiesen. Das gilt sowohl für die eingepflegten Datensets, die von allem möglichen unerwünschten Content gereinigt werden müssen (etwa Beschreibungen von Gewalt und Missbrauch, Selbstmord, Trolling und Bombenbauanleitungen). Das gilt vor allem für die späteren Fine-Tuning-Phasen, wo mit enorm aufwendigen „reinforcement learning from human feedback“ (RLHF) gearbeitet wird. Hier entsteht eine globale Arbeitsteilung. Für ersteres beauftragte OpenAI etwa Sama, eine Firma in San Francisco, die aber ihre Mitarbeiter*innen in Kenia, Uganda und Indien rekrutiert (Perrigo 2023a).

Die dort angestellten Clickworker*innen verdienen zwischen 1,30 und 2,00 Dollar pro Stunde. Die psychischen Belastungen, die durch die teils sehr drastischen Schilderungen in den zu filternden Texten entstehen, werden von den AI-Firmen kaum adressiert. In einigen Ländern gibt es bereits erste Bemühungen, sich gegen die Arbeitsbedingungen zu organisieren (Perrigo 2023b).

Die RLHF-Tätigkeit wird dagegen meist aus den USA selbst erledigt und wird entsprechen höher entlohnt. Von 10 bis 25 Dollar die Stunde ist die Rede (Dzieza 2023). Dort wird dann direkt mit dem jeweiligen Modell

interagiert, und die Antworten der KI werden bewertet. Manche Bewertungsarbeit, die eine fachliche Qualifikation erfordert, etwa im juristischen Bereich, wird teils sogar mit bis zu 50 Dollar und mehr bezahlt.

OpenAI hat derweil seine Zusammenarbeit mit Sama beendet, und insgesamt scheint sich der Annotationsmarkt aus Kenia zurückzuziehen und auf andere Niedriglohnländer zu verteilen (Perrigo 2023a; Dzieza 2023). Auch die anspruchsvollen Tätigkeiten verändern sich enorm. Je besser die Modelle werden, desto höher werden die Ansprüche an die Trainingsdaten und desto anspruchsvoller wird auch die RLHF-Arbeit.

Gleichzeitig fließt immer mehr KI-Output in die Annotationsarbeit ein. Der OpenAI-Forscher John Schulman meint, dass es immer schwieriger wird, mit den Modellen selbst Schritt zu halten (Dzieza 2023). Anthropic beispielsweise hat für Claude 2 beim Fine-Tuning teilweise schon mit automatischem Reinforcement Learning gearbeitet (Roose 2023a).

Doch der Einsatz von LLMs zur Annotationsarbeit ist nicht immer von den KI-Firmen selbst beauftragt. Auch Annotator*innen im herkömmlichen Kontext greifen heimlich auf Systeme wie ChatGPT zurück, weil es ihnen die Arbeit erleichtert. Für eine Studie zu dieser Frage ließen Wissenschaftler*innen eine Text-Zusammenfassungs-Aufgabe von Amazons Mechanical Turk erledigen und untersuchten die Ergebnisse hinsichtlich der Frage, ob und in welchem Umfang die Ergebnisse von LLMs stammen, und schätzten, dass 33 bis 46 Prozent der Zusammenfassungen KI-generiert sind (Veselovsky/Ribeiro/West 2023).

Das ist ein Teilproblem eines viel größeren Problems: Die zunehmende Verschmutzung des Internets durch den Output der LLMs selbst. Je länger LLMs frei verfügbar für die Menschen sind, desto mehr werden die neu veröffentlichten Texte selbst unter Zuhilfenahme von LLMs erstellt worden sein, die dann wieder als Trainingsdaten in neue LLMs fließen.

In ihrer Studie „The Curse of Recursion. Training on Generated Data Makes Models Forget“ haben Wissenschaftler*innen untersucht, wie LLMs wie GPT-4, aber auch Bildgeneratoren wie Stable Diffusion darauf reagieren, wenn sie wiederum mit Output von generativen KIs trainiert werden (Shumailov et al. 2023). Sie stellen fest, dass in diesem Fall ein sogenannter „model collapse“ stattfindet, ein degenerativer Prozess, bei dem die Modelle über die Zeit die zugrundeliegende Datenverteilung „vergessen“, was zu immer schlechteren Ergebnissen führt.

Eine Prognose, die sich daraus ergibt, ist, dass der Zugang zu Trainingsdaten nicht nur ein weiterer Flaschenhals ist, sondern ein Flaschenhals, der sich bereits zu schließen beginnt. Natürlich werden die Entwicklungen trotzdem weitergehen, die erstellten Trainingsdaten-Sets werden genauer geprüft werden und gleichzeitig werden sie kostbarer. Dazu wird der Einfluss von Computerpower und die Entwicklung neuerer, effiziente-

rer Trainingstechniken eine immer größere Rolle spielen und den Markt weiter konzentrieren.

Einige Beobachter wie Nick Srnicek sehen deswegen neben der Konzentration auf Ebene des Marktes auch zukünftig eine geopolitische Konzentration von AI-Entwicklung in den USA und China (Srnicek 2022). Hier setzten aktuell viel diskutierte geopolitische Überlegungen zu KI an, die hier allerdings den Rahmen dieser Studie sprechen würden.[9]

Emily Bender und Alex Hanna befürchten zudem eine neue Zwei-Klassengesellschaft auf uns zukommen. Da die generativen KIs zwar nicht gut genug sind, um menschliche Arbeit zu ersetzen, aber sich viele Leute menschliche Arbeit sowieso nicht leisten können, würden die sozial Schwachen zukünftig mit KIs abgespeist werden, während die, die es sich leisten können, sich weiterhin von Menschen bedienen, behandeln und beraten lassen (Hanna/Bender 2023).

Evgeny Morozov hat die These aufgestellt, dass KI, insbesondere der wiederkehrende Rekurs auf AGI, eine Art Radikalisierung des neoliberalen Paradigmas sei (Morozov 2023). Der Neoliberalismus verklärte die ökonomischen Verhältnisse als (gott-/markt-)gegeben und bot dem Individuum immer nur „Lösungen“ an, sich in diesen Verhältnissen zu behaupten. „Solutionismus“ sei demnach integraler Bestandteil des Neoliberalismus.

Der neuere Hype um KI und die Erzählung von der allmächtigen AGI sattelt auf diese Erzählung auf. Statt die Verhältnisse zu ändern, wird die Gesellschaft einer künstlichen Hyperintelligenz unterstellt, die Armut, Klimawandel und Kriege „löst“.

Dabei sei es unerheblich, so Morozov, ob AGI überhaupt jemals eintritt und ihre Lösungen tatsächlich funktionieren würden. Alleine die Erzählung davon hält politische Ambitionen der Gesellschaftsveränderung in Schach, ganz ähnlich wie die Politiker*innen, die gegen klimapolitische Maßnahmen argumentieren, weil die noch nicht existierende Technologie X bestimmt eh bald Abhilfe schaffen werde. KI wirkt demnach als ein „Diskurs-Device“, das hilft, politische Lösungen für die drängenden Probleme der Welt zu verhindern.

4.2 Automatisierung, Computerisierung, KI und der Arbeitsmarkt

Die Frage, wie Automatisierungstechnologien auf den Arbeitsmarkt wirken, ist so alt, dass sich bereits Karl Marx damit beschäftigte. Schon Marx

9 Als Einstieg zum Thema Geopolitik und KI empfiehlt sich Larson 2022.

spricht von Produktivitätszuwächsen durch Technologieeinsatz, die zwar von den Kapitalist*innen abgeschöpft werden, aber nur so lange bis die Konkurrenz ebenfalls die Produktivität erhöht und die Preise sinken. Marx schloss daraus den tendenziellen Fall der Profitrate (Marx 1980), neoklassische Ökonomen sehen darin den Anreiz für mehr Konsum und daher Wirtschaftswachstum. Maynard Keynes schloss aus dem stetigen Produktivitätszuwachs, dass seine Enkel wahrscheinlich nur noch drei Stunden am Tag würden arbeiten müssen (Keynes 1932).

Vor allem dem Computer als universelle Maschine wird seit seiner Erfindung die Fähigkeit zugesprochen, jede erdenkliche Arbeit zu automatisieren (Seemann 2014). In einem offenen Brief an den damaligen US-Präsidenten Lyndon B. Johnson brachten 1964 einige Wissenschaftler und Prominente eine von ihnen ausgemachte dreifache Revolution zur Sprache (The Ad Hoc Committee on the Triple Revolution 1964). Damit meinten sie einerseits die Ausweitung der Menschenrechte, wie sie unter anderem durch die Bürgerrechtsbewegung in den USA vorangebracht wurde, die Abschaffung des Krieges als Methode, internationale Konflikte zu lösen, angesichts der Aufnahme von Atomwaffen ins Arsenal der USA und der Sowjetunion und drittens etwas, das sie „Cybernation“ nannten.

Gemeint war die Computerrevolution, die, so die Autor*innen, dazu führen würde, dass nicht nur eine Menge Menschen ihre Arbeit verlieren, sondern auch, dass der Wohlstandszuwachs sich insgesamt von der geleisteten Arbeit abkoppeln würde.

Während der große Anstieg in den Arbeitslosenzahlen ausblieb, kann man festhalten, dass sich die Prognose über die wachsende Ungleichheit bestätigt hat. Nicht nur haben wir nach Jahrzehnten Computerisierung nahezu Vollbeschäftigung, auch die Produktivität ist mit der Digitalisierung nicht außergewöhnlich stark gestiegen. „You can see the computer age everywhere but in the productivity statistics“, stellte der Ökonom Robert Solow 1987 fest (zitiert nach Dudley 2014). Der Wirtschaftshistoriker Robert J. Gordon zeigte sogar, dass das Ende der großen Wachstumsperiode der Industrienationen – in den USA nach Gordon die Zeit von 1870 bis 1970 – mit der Computerrevolution zusammenfiel (Gordon 2016).

Gordon macht ausgerechnet einen Mangel an echten Innovationen in der digitalen Ära dafür verantwortlich. Und in der Tat: Vergleicht man, wie radikal sich das menschliche Leben etwa zwischen 1900 und 1940 veränderte – Elektrifizierung, Sanitäranlagen, Telefon, Automobil – mit den Veränderungen heute, bekommt man den Eindruck unser Leben habe sich seit 40 Jahren nicht wesentlich verändert. Nur die Anzahl und Größe der Bildschirme hat sich dramatisch verändert.

Das hält aber bis heute niemanden davon ab, weiterhin die radikalen Verwerfungen durch die Digitalisierung zu behaupten. Seit Jeremy Rifkin

(1995) das Ende der Arbeit ausgerufen hatte, werden solche Prognosen bei jedem Digitalisierungsschritt aufs Neue proklamiert, was von Ökonom*innen meist postwendend relativiert wird. Sie verweisen in diesem Zusammenhang gerne auf die sogenannte „lump of labor fallacy“. Diese wurde zuerst im 19. Jahrhundert vom Ökonomen David Frederick Schloss formuliert (Schloss 1891) und bezeichnet den menschlichen Fehlschluss, die Angebotsseite des Arbeitsmarktes als die Gesamtheit aller zu erledigenden Tätigkeiten zu betrachten.

Wenn man das tut, ergibt sich ein endliches Reservoir an zu erledigender Arbeit, das sich – sobald die Automatisierung einsetzt – entsprechend verringert. Es spielt in diesem Denken keine Rolle, ob die Arbeit von einem Menschen oder einer Maschine erledigt wird, die darum in einer Art Nullsummenspiel konkurrieren.

Doch so funktioniert es nicht. Der Arbeitsmarkt ist viel komplexer. Die neoklassische ökomische Sicht, wie sie z. B. von dem Ökonomen David H. Autor vertreten wird, sieht in der Automatisierung vor allem die Einsparung von Arbeitskosten, die es den Herstellern erlaubt, die Produkte günstiger anzubieten, was dann wiederum den Absatz steigert und die Produktion erhöht, was wiederum neue Arbeitsplätze schafft (Autor, 2015). Am Ende kann ein Automatisierungsschritt genauso viele Stellen schaffen wie eingespart werden.

Meist, so die Ökonomen, entstünden zudem bessere Arbeitsstellen, die mehr Qualifizierung erfordern, aber auch besser bezahlt würden. Oft sind die Prozesse sogar noch komplexer. Ein Automatisierungsschritt kann dazu führen, dass sich der ganze Absatzmarkt und damit die Struktur der Branche und aller darin enthaltenen Berufsbilder verändert. Durch die Automatisierung der Mediendistribution mittels des Internets haben wir eine komplett andere Medienlandschaft bekommen, in der es nun Berufe gibt, die sich vor 20 Jahren niemand hätte ausdenken können.

Jedoch wird diesem Mechanismus seit dem Siegeszug von KI-Systemen nicht mehr allumfassend vertraut. In dem populären Buch „The Second Machine Age“ erklären die Autoren bereits 2014, dass mit den neueren KI-Systemen eben nicht mehr nur manuelle, physische Arbeit, sondern zunehmend auch kognitive Arbeit bzw. die kognitiven Aspekte von Arbeit durch Automatisierung bedroht sind (Brynjolfsson/McAfee 2014). Die Generierung neuer Jobs bzw. die Verschiebung von Jobs auf die kognitiv nächsthöhere Ebene könnte so an eine endgültige Grenze kommen. Nämlich dann, wenn die Maschine uns in immer mehr Fähigkeiten überlegen ist.

In einer Studie untersuchten Carl Benedict Frey und Michael A. Osborne, wie sich die wachsenden Fähigkeiten von Computern auf den Arbeitsmarkt auswirken werden (Frey/Osborne 2016). Dafür griffen sie auf

eine Datenbank, das „Occupational Information Network“ (O*NET) zurück, das für allerlei Berufe die dafür notwendigen Fähigkeiten und Fertigkeiten ausweist. Für die aufgelisteten Fähigkeiten erstellten die Forscher mithilfe von Expert*innen-Befragungen Ratings, die ausweisen sollen, wie gut oder schlecht Computer diese bereits beherrschen. Auf dieser Datengrundlage schätzten sie dann ein, wie wahrscheinlich es ist, dass der jeweilige Job durch Computerautomatisierung betroffen sein wird.

Das erstaunliche Ergebnis löste ein großflächiges Echo aus: 47 % aller Jobs in den USA seien bis 2033 in Gefahr. Die Studie und die erzeugte mediale Aufmerksamkeit führten zu immer neuen Studien und Berichten. Während die OECD 2016 mit einer ähnlichen Methodik, die allerdings die Heterogenität von Berufen in unterschiedlichen Ländern mit einbezieht, nur neun Prozent Jobverlust in ihren Mitgliederstaaten sah (Arntz et al. 2016), sprach McKinsey 2017 schon von 400 bis 800 Millionen Jobs, die weltweit bis 2030 in Gefahr seien (Manyika et al. 2017).

Besonders im Transportsektor wurde vor einem regelrechten Kahlschlag gewarnt. Die Fortschritte im autonomen Fahren, die unter anderem von Google präsentiert wurden, hatten selbst Expert*innen aufgeschreckt. Viele hielten es nur noch für eine Frage der Zeit, dass Robo-Taxis und selbstlenkende LKWs Hunderttausende Jobs ersetzen würden.

In der 2022er Studie „How to compete with robots by assessing job automation risks and resilient alternatives“ greifen Paolillo et al. wiederum auf die O*NET-Datenbank zurück und verknüpften sie mit einer anderen Datenquelle, dem „European H2020 Robotics MAR“, das ein Verzeichnis über aktuelle Fähigkeiten von Robotern und KIs bereit hält (Paolillo et al. 2022). Durch das Zusammenbringen dieser beiden Datenquellen konnten die Forscher*innen für jede Berufsgruppe einen sogenannten ARI-Wert berechnen, der die Wahrscheinlichkeit eines Berufes aufzeigt, von Automatisierung betroffen zu sein.

So finden sich auf dem einen Ende des Spektrums Physiker*innen wieder, die nur schwer mit KI zu ersetzen sind und am anderen Ende der Skala Fleischverpacker*innen, die bereits heute vielfach von Automatisierung bedroht sind.

Es ist vielfach Kritik an diesen und ähnlichen Studien vorgetragen worden. Nicht nur, haben sie sich in ihren Prognosekapazitäten als wenig verlässlich erwiesen, auch ihre Methodik ist fraglich. Insbesondere das schematische Auswerten der O*NET Daten werde der Komplexität, Heterogenität und Idiosynkrasie tatsächlicher Berufsalltage nicht gerecht (Benanav 2023). Allein der Beruf des Lehrers sei innerhalb der USA und noch mehr international extrem heterogen und erfordere teils sehr unterschiedliche Fähigkeiten.

Wahrscheinlich liegt der Fehler bereits darin, Berufsbilder als Ansammlung von Kompetenz- und Fähigkeitsprofilen zu denken. Man könnte diesen Fehler die „lump of competencies fallacy“ nennen. Tatsächlich sind Berufsfelder fluide und vielgestaltig und passen sich ständig an neue Gegebenheiten an. Der Beruf des Kraftfahrers hat sich die letzten 40 Jahre auch ohne KI ständig gewandelt und umfasst heute viel mehr Management-, Logistik- und andere klassische „Schreibtisch“-Aufgaben. Dabei geht es auch immer darum, in Ausnahmesituationen flexibel und schnell reagieren zu können, was sich nur schwer mit einem automatischen Fahrassistenten erledigen lässt.

Der bereits mehrfach erwähnte „Vater der KI“, Geoffrey Hinton, sagte noch 2016 voraus, dass sich der Beruf des Radiologen in fünf Jahren erledigt haben wird. Hintergrund waren KI-Systeme, die besser als menschliche Spezialist*innen Computertomografien interpretieren konnten (Creative Destruction Lab 2016). Den Beruf der Radiolog*innen gibt es sechs Jahre später immer noch und sie werden genauso wie Kraftwagenfahrer*innen händeringend gesucht.

4.3 Auswirkungen von Large Language Models auf den Arbeitsmarkt: Die Studienlage

Aber vielleicht ist es diesmal alles anders? Im April diesen Jahres entließ die Online-Newsseite Buzzfeed einen Großteil ihrer Redakteure, kurz nachdem der Geschäftsführer Jonah Peretti gegenüber seinen Mitarbeiter*innen per Rundmail ankündigte „KI-Erweiterung in jeden Schritt ihres Verkaufsprozesses“ zu integrieren (Ropek 2023). Ähnliches passierte bei dem Tech-News-Seite CNET, die nach der heimlichen Einführung von KI im Redaktionsprozess zehn Prozent ihrer Mitarbeiter*innen freisetzte. In Deutschland will der Axel-Springer-Konzern Vorreiter beim journalistischen Einsatz von KI sein und kündigte eine „Verschlankung“ der Belegschaft der Bild-Zeitung an (Ziegener 2023).

Die Washington Post berichtete im Juni, dass einige Werbetexter*innen wegen der Aussicht auf LLM-generierte Kampagnen entlassen wurden und sich nun als Hundesitter verdienen müssen (Verma/De Vynck 2023). In China werden derweil eine ganze Menge Illustrator*innen aus dem Gaming Markt durch generative KIs verdrängt (Zhou 2023). Der einer Umfrage der Jobbörse Resumebuilder zufolge setzen laut deren Geschäftsführer*innen 49 Prozent der befragten Unternehmen Chatbots bereits ein und 30 Prozent planen es zumindest (Urban 2023). Die Unter-

nehmen sprechen mehrheitlich von den enormen Einsparungen, die sich dadurch erzielen lassen.

Während sich solche Meldungen häufen, ist die Studienlage zu der Frage, wie LLMs Jobs ersetzten, noch recht dünn. Den Autor*innen der wenigen aktuellen Studien ist durchaus bewusst, dass es noch zu früh ist, eine abschließende Einschätzung abzugeben. Viele der Studien zu LLMs greifen überdies ebenfalls auf die O*NET Datenbank zurück.

In ihrer Studie „An Early Look at the Labor Market Impact Potenzial of Large Language Models" haben OpenAI-Forscher*innen mit Wissenschaftler*innen der Universität Pennsylvania die O*NET-Kategorisierungen von Tätigkeiten innerhalb verschiedener Berufsbilder mit den Fähigkeiten von GPT-4 abgeglichen (Eloundou et al. 2023). Die Forscher*innen stellen dabei heraus, dass anders als bei bisherigen Untersuchungen zu den Effekten von Automatisierungstechnologien LLMs wie GPT-4 „General-Purpose"-Technologien seien, also keine spezialisierten KIs, um bestimmte Tätigkeiten zu automatisieren, sondern generisch genug sind, um für viele unterschiedliche Tätigkeiten infrage zu kommen.

Um die tatsächlichen Fähigkeiten von GPT-4 einzuschätzen, kamen einerseits Expert*innen-Befragungen zum Tragen, aber auch GPT-4 sollte sich selbst einschätzen. Laut diesen Daten sind 80 Prozent aller Jobs zumindest potenziell von LLMs betroffen, wobei „betroffen" keine Ersetzbarkeit per se heißen soll, sondern nur, dass in diesem Berufsfeld ein Einsatz von GPT-4 grundsätzlich vorstellbar ist. Ein weiteres Ergebnis ist, dass 19 Prozent aller Jobs zu mehr als 50 Prozent Tätigkeiten enthalten, die grundsätzlich durch LLMs erledigt werden könnten. Schließt man verwandte Technologien (andere generative KIs wie Bild-, Video-, Audio-Erzeugung) in die Betrachtung ein, sind es sogar 49 Prozent aller Jobs.

In ihrer Studie „Occupational Heterogeneity in Exposure to Generative AI" wiesen Forscher*innen KI-Anwendungen 52 menschliche Fähigkeitsbereiche zu, die sie dann wieder den Tätigkeiten der O*NET-Datenbank zuordneten (Felten/Raj/Seamans 2023). So konnten sie für jeden Beruf einen Betroffenheitswert errechnen. Interessanterweise weist die Studie gleich nach Telemarketing-Arbeitenden vor allem Lehrberufe als besonders gefährdet aus. Insgesamt warnen die Forscher*innen, dass viele der besser bezahlten Wissensarbeiter*innen-Jobs gefährdet seien. Es ergebe sich sogar eine positive Korrelation zwischen Hochschulbildung und Jobgefährdung.

In einer Studie von Goldman Sachs wird vorhergesagt, dass etwa ein Drittel aller Jobs in den USA und Europa durch KI-Automatisierung betroffen sein werden und bis zu einem Viertel der Stellen einer Automatisierung zum Opfer fallen könnte – mit den größten Automatisierungspotenzialen bei Verwaltung und Recht (Hatzius et al. 2023).

Die Bank zeichnet aber dennoch ein optimistisches Bild, da sie gleichzeitig viele Anzeichen dafür sieht, dass KI zu einem Wachstumsmotor werden kann. Sie schätzt, dass generative KI das Produktivitätswachstum in den USA um 1,5 Prozentpunkte pro Jahr anheben und für sieben Prozent der globalen Wirtschaftsleistung verantwortlich sein könnte – ihrer Ansicht nach im Rahmen dessen, wie die Erfindung des Autos und des PCs auf die Wirtschaft eingewirkt hatten.

Der „Global Jobs Report" des World Economic Forum (WEF), der eine Auswahl an internationalen Firmen zum Arbeitsmarkt befragt, zeichnet ein ähnliches Bild. Zwar glauben viele Unternehmen, dass der Einsatz von LLMs Jobs kosten wird (22 %), aber die Mehrheit (26 %) glaubt, dass es mehr Jobs schaffen wird (Di Battista et al. 2023).

Die allgemeine Theorie, wie LLMs den Arbeitsmarkt beeinflussen, ist, dass sie den Arbeitenden Produktivitätszuwächse bescheren, die wiederum in Umstrukturierungen, schlimmstenfalls in Arbeitsplatzkürzungen münden können. Das heißt, eine Texterin macht z. B. mit ChatGPT die Arbeit von zweien und deswegen muss der zweite Texter gehen. Diese Produktivitätszuwächse sind aber immer erst viel später und aggregiert in den Umsätzen der Unternehmen messbar, denn Produktivität wird immer als Umsatz geteilt durch die eingesetzte Arbeitskraft gemessen.

Eine Studie hat sich der Produktivitätsfrage experimentell angenähert, indem sie zwei Gruppen von Menschen typische Schreibtätigkeiten erledigen lies (Noy/Zhang 2023). Beide Gruppen erledigten die Tasks zunächst ohne maschinelle Hilfe. In einem zweiten Schritt wurde der einen Gruppe dann ChatGPT als Tool vorgestellt und empfohlen, der anderen Gruppe nicht. Die Gruppe, die ChatGPT einsetzte, wurde durch das Tool um 37 % Prozent schneller als die Kontrollgruppe.

Dabei veränderte sich auch die Struktur des Schreibvorgangs. Während vor ChatGPT etwa 25 Prozent der Arbeitszeit auf das Brainstorming, 50 Prozent auf das Schreiben und wiederum 25 Prozent auf die Korrektur entfiel, reduzierte sich mit ChatGPT die eigentliche Schreibzeit um etwa die Hälfte, während sich der Zeitaufwand der Korrektur verdoppelte. Die Forscher*innen versuchten auch die Frage zu klären, ob LLMs eher dazu führen, dass Jobs ersetzt werden, oder ob sie vielmehr die Fähigkeiten der Arbeitenden erweitern. Dazu schauten sie sich die Bewertung der Arbeitsergebnisse der beiden Gruppen an und befragten die Probanden, wie sehr sie ihre Arbeit durch ChatGPT gefährdet sehen.

Die Ergebnisse weisen darauf hin, dass LLMs tatsächlich eher dazu tendieren, menschliche Arbeitsleistung zu substituieren, als sie zu erweitern.

GitHub, die wichtigste Plattform für Software-Entwickler*innen, hat mit Copilot selbst ein LLM entwickelt, das speziell beim Schreiben von Code

unterstützen soll. In einer Untersuchung wollte GitHub herausfinden, wie sehr Copilot die Produktivität der Entwickler*innen beeinflusst (Kalliamvakou 2022). Produktivität bei Programmierer*innen zu messen ist besonders herausfordernd. Die Menge an produziertem Code ist zumindest ein irreführender Indikator, da z. B. die Qualitätssteigerung von Code oft gerade in seiner Verschlankung liegt.

Daher arbeitete GitHub mit zwei anderen Metriken. Einerseits eine Befragung der Userbasis: Etwa 2000 Entwickler*innen nahmen an der Umfrage teil und 88 Prozent gaben an, einen Produktivitätsschub durch Copilot erfahren zu haben. Im Detail half Copilot den Entwickler*innen schneller, sich wiederholende Aufgaben abarbeiten zu können (96 Prozent), was ihnen unter anderem erlaubte, sich erfüllenderen Aufgaben zuzuwenden (74 Prozent).

Andererseits wurden für die Untersuchung auch kontrollierte Experimente durchgeführt. Dafür wurden 95 Entwickler*innen angeworben, die in zwei Gruppen unterteilt wurden: 45 Entwickler*innen, die Copilot nutzten, und 50, die es nicht nutzen. Ihnen wurde die Aufgabe gegeben, einen Webserver in JavaScript zu schreiben, eine mittelschwere Aufgabe, die für halbwegs erfahrene Entwickler*innen auf jeden Fall lösbar ist. 78 Prozent der Entwickler*innen mit Copilot und 70 Prozent der Entwickler*innen ohne führten die Aufgabe erfolgreich aus, aber während letztere im Schnitt 161 Minuten brauchten, brauchten erstere weniger als die Hälfte der Zeit, nämlich nur 71 Minuten.

Auch Annotationsjobs ließen sich experimentell durch LLMs ersetzen. Forscher*innen der Universität Zürich ließ Menschen und LLMs gegeneinander antreten, um Annotationsaufgaben zu erledigen (Gilardi/Alizadeh/Kubli 2023). Die Forscher*innen ließen 2.382 Tweets sowohl jeweils von auf Amazons Mechanical Turk rekrutierten Clickworker*innen, speziell eingewiesenen wissenschaftlichen Mitarbeiter*innen und von ChatGPT hinsichtlich einiger Kategorisierungen annotieren. Dabei schlug ChatGPT in vier von fünf Zuordnungsbereichen die menschlichen Annotatoren von Mechanical Turk hinsichtlich Akkuratheit, die anhand der Ergebnisse der eingewiesenen wissenschaftlichen Mitarbeiter*innen bewertet wurde.

In einer anderen Metrik, dem „Intercoder Agreement“, in der nicht die Annotation einer Gruppe den Maßstab bildet, sondern die Überschneidungen der Annotationen der jeweiligen Gruppen den Ausschlag gibt, schlägt ChatGPT sogar die geschulten wissenschaftlichen Mitarbeiter*innen.

All diese Studien, Umfragen und Experimente geben Anlass zur Beunruhigung. Doch die Geschichte zeigt zuverlässig, dass die Wirklichkeit komplexer als O*NET Datenbanken, Modelle oder kontrollierte Versuchs-

anordnungen ist, weswegen ich dazu neige, diese Ergebnisse mit größter Zurückhaltung zu betrachten. Technologische Entwicklungen passieren nicht linear, sondern chaotisch, auf vielen Ebenen gleichzeitig und unvorhersehbar. Deswegen lohnt es sich, zu erwartende Metaeffekte und Strukturänderungen mit in die Betrachtung einzubeziehen.

4.4 Metaeffekte und Strukturänderungen

Das erste reale Auftauchen von LLMs auf den Arbeitsmarkt war ein indirektes. Zwar war der Streik der „Writers Guild of America“ im Sommer 2023 nicht ausschließlich, aber doch auch sehr sichtbar den neusten Entwicklungen auf dem Feld generativer KI gewidmet (Beckett 2023). Neben den extremen Lohngefällen zwischen klassischem TV/Filmproduktion und neueren online Streaming-Diensten war die Befürchtung der streikenden Autor*innen, LLMs könnten in den Skript-Produktionsprozess Einzug halten, ein wichtiger Motivator.

Die Befürchtung war aber nicht einfach, dass Hollywood sich seine Skripte von ChatGPT generieren lassen würde, sondern vielmehr, dass sich das Studio-Management von LLMs die Handlungsstruktur und das Basisskript generieren lassen und den angestellten Autor*innen dann nur noch die Überarbeitung bzw. Ausgestaltung der Details überlassen würden.

Selbst wenn das LLM nur unbrauchbaren Quatsch liefern würde, verlören die Autor*innen das Urheberrecht an dem Drehbuch und man könnte sie mit einem Bruchteil des Geldes abspeisen (Broderick 2023). Ein solches Szenario scheint die viel realistischere Blaupause für den Einfluss generativer KI auf den Arbeitsmarkt zu sein, als die vielen experimentellen Studien zur Tätigkeits-Substitution aufscheinen lassen.

Ein verwandter Effekt wird bereits länger bei Automatisierungsprozessen beobachtet. Entgegen der oft vorgetragenen Erzählung, dass sich Menschen, deren Arbeitsplatz wegautomatisiert wird, sich nach höheren Einkommensklassen orientieren, führt Automatisierung in der Realität oft dazu, dass bestimmte Arbeitsprozesse auch mit weniger Qualifizierung verrichtet werden (Rinta-Kahila 2018). Dieses sogenannte „Deskilling“ führt dann zu einer höheren Austauschbarkeit der jeweiligen Arbeitenden und in Folge zu allgemeinen Lohneinbußen in dem Bereich.

Damit spricht der Streik der Drehbuchautor*innen einen Effekt auf die Wirtschaft an, der weit einfacher nachweisbar ist: die einseitige Verteilung der Erträge des Produktivitätswachstums auf Kapitaleigner*innen statt auf Lohnempfänger*innen. Das Manifest der Triple Revolution hatte es bereits in den 1960ern bestens beschrieben:

> „Bis zu diesem Zeitpunkt wurden wirtschaftliche Ressourcen auf der Grundlage von Beiträgen zur Produktion verteilt, wobei Maschinen und Menschen unter etwa gleichen Bedingungen um Beschäftigung konkurrierten. Im sich entwickelnden cybernetischen System kann potenziell unbegrenzte Leistung durch Systeme von Maschinen erzielt werden, die wenig Zusammenarbeit von Menschen erfordern." (The Ad Hoc Committee on the Triple Revolution 1964)

Die Folge: die Abkoppelung der Wohlstandszuwächse von den Lohnzuwächsen. Genau dieser Zusammenhang zeigt sich seit den 1970ern in der stetig sinkenden Lohnquote. Über viele Jahrzehnte lag der Anteil des Bruttoinlandsprodukts, der in Form von Löhnen ausgezahlt wurde, sehr stabil bei etwa zwei Drittel. Seit den letzten 35 Jahren scheint sich das zu ändern. In Industrieländern sinkt die Lohnquote stetig: in den USA um 6 Prozent, in Deutschland um 7 Prozent, in Frankreich sogar um 14 Prozent (ILO/OECD 2015).

Eine mögliche Erklärung: Durch die Automatisierung ersetzen Kapital-Inputs Arbeits-Inputs, und die Rendite dieser Ersetzung wird einseitig der Kapitalseite zugeschlagen. Diese strukturellen Effekte sind nachhaltiger und vorhersehbarer als die Effekte auf den Arbeitsmarkt. Letztere sind komplex und vielgestaltig und äußern sich weniger im Wegfall von Arbeitsplätzen, als durch eine generelle Umstrukturierung von Branchen, Berufen oder Produktionsmodelle, die die meisten Menschen schon in ihrer Beschreibung überfordern.

Nehmen wir als Beispiel die WEF-Umfrage, in der Unternehmen zu Arbeitsplatzverlusten durch KI befragt wurden (Di Battista et al. 2023). Die Unternehmen gaben an, dass sie die derzeit von Maschinen erledigten Aufgaben bei 34 Prozent einschätzen, während der Rest, also 66 Prozent, von Menschen erledigt werden.

Doch was bedeutet das genau? Wie misst man, welche Arbeit von Menschen oder der Maschine erledigt wird? Wie viele Sekretär*innen spart z. B. die Autokorrektur von Microsoft Word ein? Wie viele Boten spart der firmeneigene E-Mailserver ein? Und sind die eingesparten Boten beritten oder schon motorisiert? Motorisiert hieße ja, dass ebenfalls Maschinen Arbeit verrichten usw. Dass diese Aussagen, wenn man sie genauer betrachtet, nur ins Absurde führen können, liegt daran, dass die Arbeitsverteilung zwischen Mensch und Maschine nach wie vor als Nullsummenspiel gedacht wird.

Umstrukturierungen ersetzten nie einfach eine Arbeitskraft durch eine KI. Nehmen wir die Vorhersage von Geoffrey Hinton, dass es in fünf Jahren keine Radiologen mehr geben werde (Creative Destruction Lab 2016). Stattdessen werden zwar KIs zur Erkennung von Krebs eingesetzt, aber nur wenn das Bildmaterial bestimmten Kriterien entspricht, z. B. mit einer bestimmten Apparatur, in bestimmten Winkel und einer richtigen Belich-

tung aufgenommen wurde. Dafür müssen dann wiederum Menschen sorgen, bevor das Bildmaterial an die KI geht, was dann wiederum von einer anderen Person geprüft wird usw. (Dzieza 2023).

Auch die Frage der Produktivitätszuwächse lässt sich nicht so leicht klären, wie die experimentellen Studien glauben machen. Ezra Klein weist in einem Meinungsstück für die New York Times darauf hin, dass es eben nicht ausreicht, individuelle Produktivität in den Blick zu nehmen.

Social Media z. B. hat auf der individuellen Ebene zwar die Informationssuche verbessert und die Möglichkeiten, mit vielen Menschen in Kontakt zu treten, enorm erweitert, aber es hat gleichzeitig unsere Aufmerksamkeit fragmentiert, uns in unsinnige Debatten verwickelt und uns auf viele andere Arten abgelenkt, sodass die negativen Effekte die individuellen Zuwächse an Produktivität aufgefressen haben (Klein 2023). Etwas Ähnliches sei für generative KIs zu erwarten. Die individuellen Produktivitätszuwächse stehen mit ziemlich großer Sicherheit neuen Produktivitätshindernissen gegenüber, die sich erst nach und nach zeigen werden.

Auf der anderen Seite werden LLMs, insbesondere, wenn sie in Office- und Kommunikations-Software integriert sind, völlig neue Möglichkeiten der Mitarbeiterüberwachung bieten. Solche KI-gestützten Überwachungspraktiken sind nicht neu, doch wenn die Software nicht nur Tastaturanschläge pro Minute misst, sondern auch erkennt, wie sorgfältig man E-Mails formuliert oder wie viele Ideen man pro Textseite produziert, erkennt, wie oft man Textbausteine verwendet und eine Gesamtbewertung des eigenen Schreibstils anbietet, dann gibt das Unternehmen völlig neue Vergleichbarkeiten, die dafür eingesetzt werden können, den unternehmensinternen Konkurrenzdruck zu erhöhen.

Und damit sind wir bei den Metaeffekten, also Effekte zweiter Ordnung, die sich aus Effekten erster Ordnung ergeben. Wenn z. B. LLMs sehr günstig Text produzieren können, wird das zu mehr Text in der Welt führen (Effekt erster Ordnung). Diese große neue Menge an Text, hat aber ihrerseits Effekte auf die Gesellschaft (Effekt zweiter Ordnung bzw. Metaeffekt).

Bedenkt man, dass laut experimentellen Studien ein Teil der Produktivitätsgewinne der LLMs bei der Text-Produktion durch zusätzliche Korrekturarbeit wieder aufgefressen wird (Noy/Zhang 2023), ergibt sich daraus ein struktureller, ökonomischer Vorteil für Produzent*innen aller Arten von Spam und Desinformation. Akteure, die es in Kauf nehmen oder gar darauf abgesehen haben, möglichst viel Schaden anzurichten, sind schlicht viel weniger darauf angewiesen, fehlerfreie Texte zu verwenden.

Während also die individuelle Produktivität durch LLMs steigen mag, wird sie vielleicht dadurch zunichte gemacht, dass alle Arbeitenden ihre Arbeitszeit dafür aufwenden müssen, sich gegen neue Formen von

Spam, Phishing und Desinformationen zu wehren und sich durch immer ausufernden Informationsmüll zu wühlen.

Ein weiterer Metaeffekt könnte sein, dass die Integrität von Kommunikation an sich untergraben wird. Schon jetzt wird von vielen Lehrenden in Schulen und Universitäten damit gerechnet, dass ein Aufsatz oder eine Hausaufgabe unter Zuhilfenahme von LLMs entstanden sein könnte (Mollick 2023c). Dieses Misstrauen hat das Zeug, das Schüler-Lehrer-Verhältnis bis an den Rand der Dysfunktionalität zu verändern.

Ähnliche Effekte sind für die gesamte Gesellschaft zu erwarten. Der Professor für Management und Innovation Ethan Mollick weist darauf hin, dass bereits jetzt LLMs in vielen Bereichen von Mitarbeitenden eingesetzt werden, ohne, dass die jeweiligen Unternehmen das wissen (Mollick 2023a). Mitarbeitende sehen die potenziellen Produktivtätsgewinne durch LLMs und versuchen sie – meist verdeckt – in ihrer Arbeit für sich zu nutzen. Mollick hat dafür den treffenden Namen „Secret Cyborgs“ gefunden. Dabei gilt es zu bedenken, dass die Heimlichkeit Teil der Wertschöpfung ist. Wie Mollick es ausdrückt: „Much of the value of AI use comes from people not knowing you are using it.“

Das führt auf die Dauer dazu, dass die professionelle Alltagskommunikation auch ohne explizite „Bad Actors“ mehr und mehr von LLM-Outputs unterwandert wird. Studien weisen konsistent darauf hin, dass Menschen computergestützte Vorschläge deutlich unkritischer akzeptieren und übernehmen, als es angemessen wäre. Ein Befund, für den sich der Name „Automation Bias“ etabliert hat (Goddard/Roudsari/Wyatt 2011). Der Effekt davon dürfte in nächster Zeit spürbar werden, wenn LLMs wie GPT-4 oder Bard in allerlei Office- und Kommunikationsanwendungen integriert werden (Mollick 2023b).

Wenn bei jeder angefangenen E-Mail oder jedem Performance Review im Word-Dokument ein Button den Schreibenden dazu verführt, den Rest des Textes in Sekundenschnelle generieren zu lassen, werden nur wenige widerstehen können.

Und hier wird ein weiterer Metaeffekt wahrscheinlich: Wie ändert sich unser Kommunikationsverhalten, wenn wir 1. ohne Mehraufwand viel, viel mehr Text produzieren und versenden können? Und wenn umgekehrt 2. jede einzelne Nachricht, die wir erhalten, pauschal unter LLM-Verdacht steht?

Es ist vielleicht nicht offensichtlich, aber der Aufwand einer Kommunikation ist immer auch Teil der Kommunikation. Dass sich jemand die Mühe macht, einen Brief zu tippen oder auch nur eine E-Mail, verleiht der Kommunikation eine gewisse Bedeutungsschwere und führt dazu, dass wir sie überhaupt ernst nehmen. Doch was passiert, wenn dieser Aufwand verschwindet oder er zumindest nicht mehr als solcher empfunden wird?

Eine konkrete Vorhersage zu machen, wie sich diese Veränderungen auswirken werden, ist an dieser Stelle unmöglich. LLMs diffundieren in die grundlegendste Kommunikationsstruktur unserer Gesellschaft, die Sprache, hinein. Das wird die Sprache an sich und unser Sprechen und Schreiben radikal verändern. Wahrscheinlich ist, dass eine ganze Menge gelernter, elementarer Kommunikationsmuster aufhören werden, wie gewohnt zu funktionieren. Unsere ganze Art und Weise, wie wir kommunizieren, wird sich daher rasant verändern.

Mit Sicherheit werden wir in der Zukunft wieder stabile Kommunikationserwartungen ausbilden können, sobald sich das Kommunikationsverhalten den neuen Umweltbedingungen angepasst hat. Aber in der Zwischenzeit ist es unwahrscheinlich, dass sich diese radikalen Einschnitte positiv auf die allgemeine Produktivität auswirken werden. Oder auf unser Leben.

Für die Zukunft der Arbeitswelt kann man festhalten, dass unterschiedliche Berufe sehr verschieden von dem Einzug von LLMs in die Arbeitswelt betroffen sein werden. Manche Berufe werden wahrscheinlich tatsächlich existenziell bedroht sein. Dazu könnten Callcenter-Angestellte, Clickworker*innen, Klatschmagazin-Redakteur*innen und Übersetzter*innen gehören. Ihnen droht die *Disruption*.

Manche Berufe werden sich hingegen kaum verändern, obwohl sich auch hier einige Rahmenbedingungen ändern und interessante Randerscheinungen eine Rolle spielen werden. Man denke z. B. an Bauarbeiter*innen, Ärzt*innen, Immobilienmakler*innen, Hausmeister*innen und Pfleger*innen. Hier ist von einer *Integration* der LLMs in den Arbeitsalltag auszugehen, ganz so, als wären LLMs nur eine neue Software im Betrieb.

Einige Berufe bleiben zwar erhalten, aber werden sich unter dem Druck der Strukturänderungen durch LLMs radikal verändern. Das dürfte die größte Gruppe sein. Zu ihr könnten alle Arten von Medienberufen zählen, sowie Lehrberufen, Verwaltungs- und Managementberufe, Beratung und vieles mehr. Sie gehen durch eine *Transformation*.

5. Die Arbeitswelt in zehn Jahren anhand von drei Beispielen

Nachdem wir erfahren haben, wie LLMs funktionieren und wozu sie in der Lage sind, und nachdem wir erörtert haben, welche Auswirkungen sie theoretisch für die Arbeitswelt haben könnten, ist es an der Zeit, konkrete Szenarien zu entwerfen, die diese Veränderungen auch vorstellbar machen.

Für jedes der drei Berufsgruppen aus den Disruptions-, Integrations- und Transformations-Szenarien wurden fiktionale Beispiele ausgearbeitet, die Ausblick auf die Arbeitswelt im Jahr 2033 geben sollen. Die (fiktive) Ex-Übersetzerin Lucia führt uns durch ihre prekäre Existenz, nachdem die LLMs von ihrem Beruf nichts übriggelassen haben. Der (fiktive) Altenpfleger Achmed zeigt uns die Vor- und Nachteile der ihn umgebenden Sprachassistenten. Und die (fiktive) Hochschuldozentin Sophie erzählt, wie sich die Universitätslandschaft in einem rasanten Jahrzehnt verwandelt hat.

5.1 Übersetzung: Das Disruptions-Szenario

Vor sechs Jahren hatte Lucia ihren letzten echten Übersetzungsjob für einen spanischen Text, also einen, den sie ohne die Zuhilfenahme einer KI erledigte. Damals waren aber schon die Preise zusammengebrochen, weswegen diese Übersetzung ein Verlustgeschäft war. Sie hatte sich lange dagegen gewehrt, mithilfe von KI zu übersetzen, aber am Ende konnte sie dem ökonomischen Druck nicht mehr Stand halten. Ihr Job hat sich seitdem zu dem einer „KI-Lektorin“ gewandelt.

Eine genaue Lektüre der Texte und ihrer maschinellen Übersetzung blieb auch nach der KI-Revolution geboten und nach wie vor produzieren LLMs Übersetzungsfehler. Meist sind es Missverständnisse bei bestimmten idiomatischen Wendungen, Mangel an Feinheiten im Sprachgefühl oder verunglückte Metaphern. Solche Fehler waren zunächst fast auf jeder zweiten Seite zu finden. Dann nur noch auf jeder Fünften, irgendwann nur noch auf jeder Zehnten. Heutige LLMs sind so gut wie perfekte Übersetzer. Leider.

Die Welt hat sich radikal gewandelt. Man spricht in der Branche – oder dem, was von ihr übrig ist – vom „ChatGPT-Moment“ oder eben von der „LLM-Revolution“. Dabei hat es für Übersetzer*innen viel früher angefangen. Schon Google Translate hatte sich schnell in den Aufträgen bemerkbar gemacht. Das war 2006, doch der Dienst war damals noch relativ

schlecht. Je besser er wurde, desto mehr Menschen trauten sich zu, Texte selbst zu übersetzen, zumindest einfache Gebrauchstexte. Ein weiterer Einschnitt war DeepL 2017. Das war die erste Maschine, die zuverlässig gut lesbare Übersetzungen bot, für zahlenden Kunden sogar für längere Texte.

Doch kein Dienst hatte so eine Breitenwirkung wie ChatGPT ab 2022. Und seit vor acht Jahren einige Anbieter von LLMs das Kontext-Fenster auf eine Million Tokens und mehr ausgelegt haben, werden fast alle Übersetzungen – ob Bücher, Nachrichtenartikel, technische Dokumentationen oder Websites – von LLMs durchgeführt. Es ist fast egal, welche Zielsprache man anpeilt, sie übersetzen präziser und zuverlässiger als die meisten Profis. Sie sind sogar in der Lage, Kontext und kulturelle Nuancen zu erkennen und genau wiederzugeben, was lange als das letzte herausragende Merkmal menschlicher Übersetzer*innen galt.

Sogar handschriftliche Dokumente werden mittlerweile mittels KI-Systemen zuverlässiger entziffert und übersetzt als durch menschliche Übersetzer*innen. Große Übersetzungsunternehmen und Dienste setzen fast ausschließlich auf KI-Modelle, und viele kleinere Anbieter und Freiberufler*innen sind aus dem Markt ausgeschieden, weil sie mit den niedrigen Kosten und den schnellen Lieferzeiten der LLMs nicht mithalten konnten.

Es gibt Ausnahmen. Lucia hat eine Bekannte, Karo, die weiterhin für Übersetzungen aus dem Französischen angefragt wird. Allerdings ist sie nebenbei selbst auch renommierte Autorin. Die Kunst- und Literaturübersetzung ist ein Bereich, in dem menschliche Übersetzer*innen immer noch wertgeschätzt werden, da sie eine einzigartige persönliche Interpretation und Sensibilität in ihre Arbeit einbringen können, die die Modelle nicht reproduzieren können. Das ist zumindest die Erzählung. Dahinter steckt aber auch ein Stück Nostalgie.

Heimlich, so gestand Karo ihr einmal beim gemeinsamen Mittagessen, verwende sie auch LLMs. „Es ist einfach so praktisch!“ Die Diskussion um „Secret Cyborgs“ ist zehn Jahre alt, viele haben sich längst zu ihrem Einsatz von LLMs bekannt. Doch gerade für Menschen mit einem gewissen Renommee ist dieser Schritt anscheinend noch zu riskant.

Für die anderen bleibt der KI-Lektorats-Job. Wie Lucia korrigieren und verbessern viele ihrer Kolleg*innen die Übersetzungen von Maschinen. Manche verdienen sich auch ein Zubrot damit, die Modelle selbst zu trainieren. Das nennt sich Fine-Tuning und hilft anscheinend ihre Genauigkeit und Natürlichkeit in der maschinellen Übersetzung weiter zu verbessern. Das sind aber immer nur kurzfristige Projekte, dafür umso besser bezahlt.

Einige Übersetzer*innen haben sich auch als „Übersetzungsberater*innen“ neu erfunden und bieten ihre Expertise in den Bereichen KI-Training und Qualitätssicherung an, doch dafür hat Lucia ein zu distanziertes Verhältnis zur Technik. Eigentlich kann man fast von einer Feindschaft sprechen. Zumindest von ihrer Seite aus.

Es gibt andere, die trauen sich sogar zu, Prompt-Übersetzungen zu machen. Filmskripts enthalten nicht mehr nur Angaben zu Szenerie und Schauspielanleitungen, sondern auch immer mal wieder Prompts für Tools zur Erzeugung von Videos. Heutige Videobearbeitungssoftware hat generative KIs für Bilder und Filmsequenzen eingebaut. Wenn im Film ein Brief oder nur eine Leuchtschrift auftaucht, kann die durch Prompt-Eingabe schnell angepasst werden.

Die Herausforderung: Man muss Details wie Schattenwurf, Schriftart und Hintergrund im Prompt genau vermerken, damit es wirklich gut wird. Meist muss man viel probieren, bis man den richtigen Prompt trifft, sprachliche und kulturelle Kontexte können dabei eine Rolle spielen. Man braucht aber vor allem ein entsprechendes technisches Verständnis, das Lucia abgeht. Sie bleibt beim traditionellen Lektorat und kann froh sein, dass sie eine der wenigen ist, die bis hierhin durchgehalten haben.

Gerade liest sie ein maschinell übersetztes Skript einer Fernsehserie. Sie ist bereits auf Seite 45, aber noch hat sie keine Anmerkung gemacht. Ab Seite 50 wird sie meist nervös. Dann blättert sie im Manuskript zurück (sie druckt sich die Seiten immer noch aus) und liest noch mal genauer. So eine Perfektion seitens der Maschine verleitet zu Spitzfindigkeiten.

Neulich rief sie ein entnervter Auftraggeber an, weil sie die Maschinenübersetzung von „estar como una cabra“ mit „verrückt sein“ nicht gelten ließ. Der Ausdruck bedeutet wörtlich „wie eine Ziege sein“, wird aber metaphorisch dafür verwendet, auf irrationales Handeln hinzuweisen. Lucia war die Übersetzung zu platt, denn sie ließ die spielerische, leicht humoristische Note der Metapher vermissen. Sie schlug stattdessen „einen Vogel haben“ vor. Sie war nach wie vor überzeugt, dass das die bessere Übersetzung war, aber der Kunde grummelte nur: „Verrückt gefällt mir besser. Und das kriege ich umsonst!“

Auch das haben LLMs geändert: die Leute rufen wieder mehr an. E-Mails, SMS, Textnachrichten per Messenger, all das nimmt niemand mehr wirklich ernst. Manche haben die entsprechenden Apps komplett gelöscht. Es war einfach viel zu viel Spam. Vor allem: wirklich überzeugender Spam. Mit persönlicher Ansprache, oft mit Details aus dem eigenen Leben gespickt. Die besonders perfiden Spam-Nachrichten hatten sogar den Absender gefälscht. Freund*innen, Bekannte, Kolleg*innen, manchmal die eigenen Kinder schrieben einem. Spamfilter wurden nutzlos. Und etwa die Hälfte der Nachrichten waren Phishing-Versuche.

Das schlimmste aber ist, dass sich selbst bei den echten Nachrichten ein allgemeines Misstrauen eingestellt hat. „Hat sie die Nachricht wirklich selbst geschrieben, oder sie generieren lassen?“ Es ist viel zu einfach geworden, seit jedes E-Mailprogramm „den Button“ hat. Dann lieber Telefon. Natürlich kann man auch Telefonanrufe fälschen, aber der Aufwand ist vergleichsweise hoch, deswegen gibt es noch ein generelles Zutrauen zur fernmündlichen Kommunikation. Im Zweifelsfall wechselt man in den Videomodus. Der Nachteil ist, die Genervtheit in der Stimme ihrer Auftraggeber hören zu müssen, aus der sie den Wunsch heraushören kann, sie als immer überflüssiger werdenden Kostenfaktor endlich loszuwerden.

Auf Seite 52 findet sie doch noch einen gravierenden Fehler. „La papa es importante en muchas culturas“ übersetzte die Maschine mit „Der Papst ist in vielen Kulturen wichtig“. Doch weil es sich um eine Dokumentation zu Ernährungsfragen in Lateinamerika handelt, weiß Lucia, dass hier „die Kartoffel“ gemeint ist, nicht der Papst. Leicht beschwingt streicht sie die Stelle an und notiert die Korrektur am Rand. Heute hat sie ihr Geld wieder verdient.

5.2 Pflege: Das Integrations-Szenario

Achmed hält die rechte Seite von Herrn Bensens Oberkörper hoch und betrachtet seinen Rücken: „Ein Hämatom am Schulterblatt, eines am Steiß.“ Er spricht langsam und deutlich. Das hat er sich antrainiert, als vor fünf Jahren die ersten Sprach-Notiz-Systeme eingeführt wurden. Die waren noch auf deutliche und langsame Aussprache angewiesen. Die jüngeren Kolleg*innen sprechen ganz normal mit der KI, und das klappt auch. Aber Angewohnheiten sind mächtig.

Manchmal wünschte sich Achmed, es wäre das eingetreten, was in seiner Ausbildungszeit als das Szenario für den Einsatz für Maschinen angenommen wurde. Damals sprach man von „Pflegerobotern“, und wenn es so gekommen wäre, dann hätte er zwar immer noch seine Notizen am Computer abtippen müssen, aber das Halten des Patienten wäre ihm abgenommen worden. Es hat sich herausgestellt, dass Patient*innen nicht gern von Robotern angefasst werden wollen und automatisch transkribierte Notizen sind ebenfalls nützlich, vor allem, wenn man keine Hand frei hat.

Serio, so heißt das hauseigene System (es ist nicht selbst entwickelt worden, aber man kann es beliebig anpassen), kann auch allerlei Fragen beantworten. Dabei erkennt es, ob Pflegekräfte, Ärzte oder Patient*innen fragen, und antwortet jeweils in einem anderen Modus.

Eigentlich wird Serio vor allem von Patient*innen genutzt, die über das System Bedarfe oder Besucher*innen anmelden, Fragen stellen oder sich einfach ein Buch vorlesen lassen. Achmed würde sich nie die Blöße geben, dem System vor Patient*innen eine Wissensfrage zu stellen. Umgekehrt konfrontieren ihn Patient*innen häufig mal mit Fragen oder Einwänden zur Behandlung, die sie bei Serio erfragt haben. Das ist einer der unangenehmeren Seiten der KI.

Eine andere unangenehme Seite ist, dass Achmed ständig das Gefühl verspürt, dass Serio ihn mit wertendem Blick über die Schulter schaut. Serio speichert, wie lange Achmed bei jedem Patienten verweilt, nimmt Notiz von Nuancen in Achmeds Ansprache, registriert jeden noch so kleinen Fehler in der Behandlung. Beim Mitarbeitergespräch liegt bei seinem Vorgesetzten dann ein detaillierter Bericht auf dem Schreibtisch, der alle Fehltritte von Achmed zusammenfassend bewertet. Achmed macht sich dennoch keine großen Sorgen, denn er weiß, dass er seinen Job gut macht. Doch dieses unangenehme Gefühl der ständigen Überwachung geht dennoch nicht weg.

Doch so grundlegend hat sich sein Beruf nicht gewandelt. Natürlich sind die Maschinen um ihn herum besser geworden. Das Pflegebett hat schon einige nützliche neue Funktionen und die vielen Sensoren, die jetzt unauffällig alle möglichen Werte am Patienten messen, helfen bei einer gezielteren und effektiveren medizinischen Versorgung. Doch die Arbeit am Körper ist weiterhin eine menschliche Arbeit, der Einsatz der Hände ist nicht nur eine Frage von mechanischer Kraft und feinmotorischer Sensibilität, sondern auch eine der Empathie, Zuwendung und des Mitgefühls: etwas, was die Maschinen nicht leisten können.

Die eingesprochenen Notizen und die Daten der Sensoren werden von der KI weiterverarbeitet und in tägliche Berichte für die Ärzteschaft zusammengefasst. Die KI hat nicht nur das gesammelte medizinische Wissen der letzten 100 Jahre gespeichert, es weist auch von sich aus auf Irregularitäten hin, listet mögliche Diagnosen auf und gibt Empfehlungen für weitere Untersuchungen oder Behandlungen. Das ist ungemein nützlich und hat in vielen Fällen dazu geführt, dass Leben gerettet wurden.

Es ist aber auch gefährlich, weil der Anreiz für die Ärzt*innen sehr groß ist, sich nur im Rahmen der Berichte zu orientieren und das selbstständige Denken zu verlernen. Das hat ebenfalls schon Leben gekostet, weil eine Krankheit wegen ungewöhnlichem Symptomverlauf von der KI nicht rechtzeitig erkannt wurde. Auf der Station haben sie seitdem eine rotierende Position, wo ein Arzt oder eine Ärztin jeweils ein halbes Jahr von dem Output des Systems abgeschnitten wird und die Patient*innen nur nach eigenen Anschauungen beurteilen muss. Die betreffende Ärztin oder der betreffende Arzt muss dann bei wichtigen Entscheidungen hin-

zugezogen werden und kann bei Behandlungsentscheidungen intervenieren.

Nachdem Achmed den Rücken des Patienten eingesalbt hat, setzt er ihn aufrecht ins Bett. „Gleich kommt das Abendessen, Herr Bensen. Ihre Stellen sind besser geworden, aber wir beobachten das weiter. Machen sie noch Krankengymnastik?“ Natürlich weiß Achmed, dass Herr Bensen Krankengymnastik macht, denn er liest sich vor jedem Besuch immer nochmal kurz das automatisch erstellte Datenblatt durch, bevor er das Zimmer betritt. Aber zu viel Wissen kann Kommunikation hemmen und Kommunikation, das weiß Achmed, ist wichtig.

Ein typischer Fehler, den jüngere Kolleg*innen machen, ist das Wissen, das sie haben, wortlos anzuwenden. Patient*innen brauchen aber das ständige Gespräch, um sich in die Behandlung eingeschlossen zu fühlen. „Aber sie haben doch längst zugestimmt …“ hilft da nicht weiter.

Überhaupt die jüngeren Kolleg*innen. Die meisten sind schon an Simulationen geschult worden. Man merkt das daran, dass sie zwar über ein umfangreiches theoretisches Wissen verfügen und ihr Umgang mit den KI-Systemen viel natürlicher wirkt, sie aber oft noch etwas ungeübt sind, wenn sie mit der unübersichtlicheren Realität konfrontiert werden. Patientensimulationen sind zwar erstaunlich gut geworden, aber es bleiben eben doch Simulationen. Wie man eine Verbindung herstellt, wie man Vertrauen erwirbt und wie man mit Grenzsituationen umgeht, all das lernt man nur an echten Patient*innen. Allen voran die Fähigkeit, den Menschen hinter all den Daten und Fakten zu erkennen.

Derweil wird das Tablett mit dem Abendessen hereingefahren. Das machen in der Tat flache, fahrende Roboter. Das Essen ist personalisiert vorbereitet worden, aus einerseits den Wünschen des Patienten, aber auch den Diätvorgaben der behandelnden Ärzt*innen. Die tatsächliche Essenauswahl und die Zubereitung ist KI-gesteuert und wird von großen, zentralisierten Dienstleistern als Service für Krankenhäuser, Kur- und Pflegereinrichtungen angeboten. Trotz der Personalisierung ist die häufigste Beschwerde, dass das Essen sehr generisch schmeckt. „Egal, ob man Fisch oder Sauerkraut serviert bekommt, es schmeckt alles nach Huhn“, pflegt Herr Bensen zu scherzen.

Natürlich sind dadurch Stellen eingespart worden. Aber Arbeitslosigkeit ist so ziemlich das Letzte, worüber sich Achmed Sorgen macht. Die Demografie in Deutschland sichert ihm seinen Job mindestens, bis er selbst in Rente geht. Tatsächlich wird überall händeringend nach Personal gesucht. Das hat sich auch bei seiner Verhandlungsposition gegenüber seinem Arbeitgeber ausgewirkt, weswegen er nun ein 14. Monatsgehalt bekommt. Auf der anderen Seite musste er oft Überstunden machen, weil die Arbeit anders nicht zu schaffen war.

Die Lage hat sich etwas entspannt, seitdem die Kenntnis der deutschen Sprache kein Einstellungskriterium mehr ist. Serio ist ein prima Simultanübersetzer. Meist dauert es aber eh nicht lang, bis die neuen Kolleg*innen ihr Deutsch verbessern. Wenn die LLMs etwas gebracht haben, dann den allgemeinen Zugang zu wirklich guten, persönlichen Sprachtrainern.

Achmed verabschiedet sich von Herrn Bensen, während ihm bereits das Datenblatt des nächsten Patienten eingeblendet wird.

5.3 Bildung und Forschung: Das Transformations-Szenario

Sophie hört konzentriert zu, wie René Welms die Ergebnisse des Bachelorprojekts vorstellt, an dem er mit anderen Kommiliton*innen in den letzten Wochen zusammengearbeitet hat. Es ist zwar eine Gruppenarbeit, dennoch werden die Studierenden jeweils einzeln dazu befragt.

Der Anteil mündlicher Prüfungen hat enorm zugenommen, seit vor zehn Jahren die „Homework-Apokalypse“ zugeschlagen hat. Mit dem Aufkommen allgemein zugänglicher LLMs wurden alle Formen unüberwachter, schriftlicher Leistungsnachweise auf einen Schlag wertlos.

Die Arbeitserleichterungen, die Systeme wie ChatGPT oder Claude anboten, waren einfach zu verführerisch, als dass selbst die fleißigsten Studierenden ihnen widerstehen konnten. Das Resultat war eine Flut von erstaunlich gut lesbaren Hausarbeiten und Essays, von denen aber klar war, dass nur ein Bruchteil davon selbstständig verfasst wurden. Einigen der Arbeiten merkte man wenigsten an, dass sie aus eigenen Gedanken und LLM-Output zusammengeflickt waren.

Sophie war damals selbst noch Studierende, arbeitete aber schon an ihrem Masterabschluss. Auch sie konnte sich nicht der Versuchung erwehren ihre Forschungsfragen in das System einzugeben. Erst war es nur Neugier. Aber was das System ausgab, war stellenweise besser, als das, was sie sich zurechtgelegt hatte. Großzügig kopierte sie die eine oder andere Textstelle in ihre Arbeit: „Wird schon keiner merken.“ Hat auch niemand gemerkt, denn damals war das Misstrauen noch nicht so groß.

Heute nimmt sie selbst Prüfungen ab. Die mündliche Prüfung hat viele andere Formate verdrängt, weil sie so schwer zu hintergehen ist. Aber auch Klausuren auf speziell bereitgestellten Schreibcomputern haben vermehrt Einzug gehalten. Manche Fachbereiche lassen Klausuren sogar wieder handschriftlich schreiben.

Sophie ist wissenschaftliche Mitarbeiterin auf einer Postdoc-Stelle an der Universität Greifswald im Studiengang Erziehungswissenschaften. Es sind nicht nur die Hausaufgaben und Prüfungsleistungen, die sich verändert haben. Die Universität ist heute ein ganz anderer Ort.

René, ihr Student, hatte mit anderen zusammen eine Studie durchgeführt, die die Auswirkungen verschiedener Lehrmethoden auf das Lernverhalten von Grundschüler*innen erforschen sollte. Er selbst war für den Methodenteil verantwortlich und so fragt Sophie ihn, warum er sich bei der Auswertung für die multivariaten Varianzanalyse entschieden habe. René kommt ins Stottern. Nach einigem Rumgedruckse wird klar: er kann nicht einmal erklären, mit welchen Annahmen dieses Verfahren arbeitet und wie er sie für seine Studie geprüft hatte. Er hatte auch keine Gedanken über die Formulierung der Fragebögen und wie sie die Ergebnisse verzerren könnten.

Solche Lücken kommen eigentlich nicht vor, wenn man selbst eine so komplexe Analyse vorbereitet und durchführt. Sophie wird klar, dass hier wieder ein LLM die Hauptarbeit gemacht hat.

Unter Kolleg*innen firmieren solche Vorkommnisse als „Ungleichzeitigkeiten“. Studierende kommen selbst im Grundstudium mit hochtrabenden, komplexen Methodenapparaten an, und ihre Anwendung sieht auf den ersten Blick makellos aus. Doch bohrt man auch nur ein bisschen an der Oberfläche, können sie die einfachsten Fragen zu ihrer Methodik nicht beantworten. Es ist, als hätte Superman fliegen gelernt, bevor er krabbeln konnte. Sophie ist genervt.

Natürlich hat sie Verständnis dafür, dass die Studierenden mit möglichst wenig Aufwand möglichst beeindruckende Ergebnisse erzielen wollen. Der Druck auf die Studierenden ist groß. Aber auch sie musste, wie alle, nach dem Aufkommen der LLMs neu lernen zu lernen. Sie musste lernen, zwischen Arbeitsergebnis und eigenem Lernprozess zu unterscheiden. Sie musste lernen, sich kritisch zu befragen: Hast du das jetzt wirklich verstanden? Sie musste lernen, das Lernen seiner selbst wegen wieder wertzuschätzen.

Klar, Sophie hatte Glück gehabt. Sie wurde direkt zur Homework-Apokalypse fertig mit dem Studium. Die Dissertation war da schon schwieriger. Sie hatte eine kumulative Promotion gemacht, bei der man mehrere Aufsätze in einschlägigen Fachjournalen publiziert, statt eine große Dissertation zu schreiben. Das erste Paper ging noch gut durch. Das zweite – das war so 2024 – rasselte direkt in die „Paper Explosion“.

„Paper Explosion“ nennt man heute das Ereignis ab Mitte der 20er-Jahre, als sich das Aufkommen eingereichter Aufsätze und Studien in den Journalen um über 50 Prozent pro Jahr erhöhte, bevor man 2026 die Reißleine zog und nichts mehr annahm. Es waren nicht nur die LLM-

generierten Quatschpaper, die das wissenschaftliche Veröffentlichungssystem verstopften. Auch die seriösen Wissenschaftler*innen wurden durch LLMs einfach wahnsinnig viel produktiver und publizierten statt ein oder zwei bis zu fünf oder sechs Paper pro Jahr.

Gleichzeitig hob sich das Qualitätsniveau der eingereichten Aufsätze hinsichtlich Schreibstil und formalen Kriterien im Schnitt sogar noch, was es Verlag und Reviewer zusätzlich erschwerte, eine Auswahl zu treffen. Die Reviewer kamen schlicht nicht mehr hinterher, mehr und mehr Publikationen kapitulierten vor den Einreichungen. Sophies zweites Paper brauchte dreieinhalb Jahre von Einreichung bis Publikation. So richtig düster sah es dann für das dritte Paper aus, als die meisten Journale bereits den Annahmestopp verkündet hatten.

Es war aber auch genau in dieser Zeit, als sich aus einem Krisenrat der Universitäten eine internationale Kommission zur Neudefinition wissenschaftlicher Standards bildete. Sophie, die selbst im Bereich der Wissenschaftstheorie forschte, verfolgte die Arbeit der Kommission sehr genau und beteiligte sich mit eigenen Debattenbeiträgen an dem Prozess. Es wurden ganz neue Systeme debattiert, es wurde leidenschaftlich gestritten und gerungen. Die Kommissionsarbeit dauerte drei Jahre und während sie tagte, stellte sich heraus, dass es nicht reichte, nur die Homework-Apokalypse und die Paper-Explosion zu adressieren.

Hinter der Paper-Explosion versteckte sich ein noch viel grundlegenderes und schwierigeres Problem: Bislang basierte ein uneingestanden großer Teil der Bewertungsarbeit für wissenschaftlichen Publikationen darin, festzustellen, dass der oder die Autor*in einen gewissen fachspezifischen Ton triff.

Es war ja nicht nur so, dass es eine gewisse fachsprachliche Wiedererkennbarkeit in wissenschaftlichen Publikationen gab, nein, jede Disziplin bildete eine ganz eigene, spezifische Sprache aus. Ein eigenes Vokabular, ein Thesaurus eigener Abkürzungen und sogar die ein oder andere eigentümliche rhetorische Figur. Diese Sprache zu beherrschen war bislang ein untrügliches Zeichen dafür, dass angehende Wissenschaftler*innen sich entsprechend lang und tief in die kanonische Literatur eingelesen haben bzw. selbst Erfahrung haben, in diesem Bereich zu publizieren. Das berühmte Peer-Review war immer schon zu einem Gutteil eine Peer-Recognition.

Doch nachdem die LLMs diese Fachsprachen perfekt beherrschen lernten, brach diese Heuristik vollständig zusammen. Man sprach von der sogenannten „Jargon-Krise“. Ein perfekt in der eigenen spezifischen Fachsprache verfasstes Paper war kein hinreichendes Indiz mehr dafür, dass man es mit einem oder einer Expert*in zu tun hat. Dieser Zusammenbruch von Erwartungsstrukturen stellte sich als die allergrößte Her-

ausforderung heraus. Aber auch als Chance, wissenschaftlichen Wert ehrlicher und direkter zu repräsentieren.

Sophies drittes Paper war eines der ersten, dass unter den neuen Maßgaben der Kommission verfasst, eingereicht und bewertet wurden. Für die Publikation wissenschaftlicher Aufsätze wurde folgende Vorgehensweise vorgeschlagen:

- Schritt 1: Alle eingereichten Paper werden durch spezielle LLMs in eine generische Standardsprache übersetzt. Diese fungiert gewissermaßen als ein zusätzlicher Schritt in der Anonymisierung, der nebenbei aber auch die unbewusste Bewertung nach „Stallgeruch“ verunmöglichen soll.
- Schritt 2: Gleichzeitig findet eine automatisierte Plausibilitätsprüfung statt, die bereits methodische oder logische Fehler oder argumentative Schwächen erkennt und die das Paper bereits mit entsprechenden Hinweisen zurückweisen kann.
- Schritt 3: Erst, wenn diese Hürde genommen wurde, bekommt der menschliche Reviewer das Paper. Er bekommt aber dazu auch eine von der KI zusammengestellte Sammlung von ähnlich gelagerten Papers, mit vorgefertigten Zusammenfassungen und Hinweisen darauf, wo sich Überschneidungen oder Widersprüche zur bisherigen Forschung vermuten ließen. Das versetzt die Reviewer in die Lage, schnell den thematischen Kontext des eingereichten Papers zu erfassen und den wissenschaftlichen Beitrag des Papers einschätzen zu können.
- Schritt 4: Erst dann werden – je nachdem ob es sinnvoll ist, wiederum unter Zuhilfenahme von LLMs – die gemachten wissenschaftlichen Beobachtungen oder Überlegungen geprüft und nachvollzogen. Das ist ein Schritt, der zwar auch bislang für jedes Peer Review unterstellt wurde, aber in der Realität viel zu oft unter den Tisch fiel oder nur sehr oberflächlich passierte. Das neue System verlangt aber genau für diesen Prozess sehr genaue Dokumentationspflichten.
- Schritt 5: Erst nach abgeschlossenem Peer-Review bekommen die Reviewer das originale Paper zu Gesicht, allerdings nur um zu bestätigen, dass es mit der von ihnen untersuchten LLM-Übersetzung inhaltlich identisch ist. Wenn nicht, muss der ganze Prozess wiederholt werden.

Dieses System ist nicht nur viel effizienter, sondern auch in vielen Gesichtspunkten gerechter und sorgt für eine allgemeine Qualitätssteigerung in den Fächern, wo es angewendet wird.

Es hat aber auch das wissenschaftliche Schreiben an sich verändert. Da der Zwang, einer bestimmten Fachsprache zu entsprechen, entfallen

ist, lesen sich die Paper im Original viel blumiger und sprachlich abwechslungsreicher, einerseits um sich habituell möglichst weit von dem Verdacht, mit einem LLM gearbeitet zu haben, zu entfernen, aber auch um von vornherein ein möglichst großes Publikum anzusprechen. Ein weiterer positiver Effekt ist, dass die Wissenschaftler*innen nicht mehr dünne Ergebnisse mit einer maximal unverständlichen „wissenschaftlichen" Sprache kaschieren können.

Da Sophies drittes und letztes Paper eines der ersten des Verfahrens war, ist es noch recht brav geschrieben. Sie schloss mit Summa Cum Laude ab und wurde gleich in den Universitätsbetrieb übergenommen.

Damals wurden gerade viele Stellen geschaffen, weil die neuen Publikationsrichtlinien dazu führten, dass auch ältere Paper noch einmal mit den neuen Prozessen reevaluiert wurden. Es herrschte eine Art Aufbruchstimmung. Ein neuer wissenschaftlicher Blick bildete sich, der weniger anfällig für sozialen und kulturellen Bias war. Nicht alles hielt diesem neuen Blick stand. Das führte natürlich zu einigen Konflikten, vor allem mit den älteren Kolleg*innen, die ihr Lebenswerk infrage gestellt sahen. Aber auch sie konnten sich dem grundlegenden Veränderungsprozess nicht erwehren, der wie eine große Welle durch alle Wissenschaftsinstitutionen der Welt schwappte.

Auch die Lehre war natürlich von den Veränderungen betroffen. Zum einen waren Vorlesungen schon lange vor ChatGPT überflüssig geworden. Youtube und die vielen Online-Universitäten hatten bereits die Infrastruktur für eine skalierbare reine Online-Lehre bereitgestellt. Mit generativer KI konnte man sich nun auch noch aussuchen, wer die Lehrperson war. Populär war, sich die Grundlagen der Pädagogik von einem Deep Fake des Schauspielers Morgan Freeman erklären zu lassen.

Zum anderen lernten die Studierenden mithilfe von LLMs sehr viel mehr allein zu Haus. Sie hatten plötzlich einen omnikompetenten Tutor zur Seite gestellt, der mit unendlicher Geduld jedes Konzept, jede Idee und jede Methode für jedes Lernniveau so oft erklären konnte, wie es eben notwendig war. Die soziale Komponente des Lernens kam entsprechend zu kurz. Lerngruppen waren selten geworden, zur Vorlesung traf man sich auch nicht mehr.

Um trotzdem studentischen Austausch zu ermöglichen und zu fördern, schlug die Kommission deswegen die Ausweitung von Seminaren und studentisch organisierte Diskussionsveranstaltungen zu kontroversen Fachthemen vor. Das wurde gern angenommen. Überhaupt erhöhten sich die studentisch organisierten Veranstaltungen dramatisch, seit sie von den Universitäten explizit gefördert wurden.

Der universitäre Mittelbau ist seitdem zwar nicht verschwunden, aber seine Tätigkeiten haben sich radikal verändert. Klassische Lehre wird nur

noch von sehr wenigen, dafür aber rhetorisch sehr begabten Kolleg*innen betrieben. Ihre Online-Vorlesungen haben jedes Jahr hunderttausende von Abrufen und sie sind universitätsübergreifend, manchmal sogar international, bekannt.

Die weniger bekannten Kolleg*innen betreuen sogenannte Kolloquien, so heißen mehr oder minder stabile Gruppen von Studierenden einer Fachrichtung, die sich selbstständig zusammengefunden haben, um zusammen durch das Studium zu navigieren.

Dabei spielt weniger der fachliche Wissenstransfer eine Rolle als vielmehr eine generelle Unterstützung und Hilfestellungen bei der Orientierung. Man ist für die Studierenden so etwas wie der Vertrauenslehrer / die Vertrauenslehrerin und der Hauptansprechpartner*in für die meisten Fragen. Es gibt regelmäßige Treffen, aber auch viel Eins-zu-eins-Austausch. Im Groben kann man sagen, dass Kolloquien den Studierenden helfen, sich selbst zu organisieren und sie mit sozialen Kontakten versorgen.

Lehrende sind nicht nur Wissensvermittler. Sie sind auch Vertrauenspersonen, Rollenvorbilder und im besten Fall Mentor*innen, die junge Leute mit Orientierung, Werten und Identitätsressourcen versorgen. Eine KI kann so etwas nicht leisten.

Sophie weist René darauf hin, dass sein LLM-Einsatz gegen die ethischen Richtlinien der neuen Prüfungsordnung verstößt und macht ihm den Vorschlag, die Bachelorarbeit zu wiederholen. Der Einsatz von LLMs zur Unterstützung der eigenen Projekte ist nicht per se unethisch, er wird sogar explizit empfohlen. Man ist aber angehalten, die zentralen Aussagen von LLMs in der Tiefe nachvollziehen zu können und ggf. mit eigenen Nachforschungen zu überprüfen, was René offensichtlich versäumt hat.

Sophie muss sich schmunzelnd daran erinnern, dass sie als Studentin selbst nicht besser war. Doch es hilft ja nichts: Erwischt ist erwischt.

6. Fazit

LLMs werden der Kern einer neuen Welle von immer umfassenderen Assistenzsystemen werden, die nach und nach in die Arbeitswelt integriert werden. Die Entwicklung geht rasant voran und innerhalb der nächsten zehn Jahre ist mit einem stetigen Anwachsen von potenziellen Einsatzzwecken zu rechnen.

Ein großes Fragezeichen bleiben dabei die Open-Source-Modelle. Es ist aus heutiger Sicht unwahrscheinlich, dass sie zu den geschlossenen, kommerziellen Systemen aufschließen. Aber selbst, wenn sie in der Entwicklung um ein Jahr hinterherhinken, können sie enorme Effekte auf die Machtverhältnisse in der Wirtschaft haben. Wenn Privatpersonen und Unternehmen zunehmend autonom agierenden Bots Aufgaben anvertrauen, werden sie nach Gewissheiten suchen, dass das Programm nicht im Verborgenen fremden Interessen dient.

Kaum ein Bereich der Arbeitswelt wird unberührt bleiben, auch wenn diese Berührung in ihrer Qualität sehr unterschiedlich ausfallen wird. Arbeitende, die vom Disruptions-Szenario betroffen sind, stehen vor der Herausforderung, sich komplett neu zu orientieren. Angehörige von Berufen, auf die das Integrations-Szenario passt, können sich zwar erst einmal sicher fühlen, doch auch hier kann der zunehmende Einsatz von KI zu Personalabbau und erhöhtem Leistungsdruck führen.

Für Beschäftigte im Transformations-Szenario werden die nächsten Jahre ein wilder Ritt, in der sie ihre neue Rolle finden und erfinden müssen. Aber weil das ein so unbestimmter, chaotischer Prozess sein wird, steckt hier auch eine Menge Gestaltungspotenzial. Für alle Bereiche aber gilt, dass die allgemeinen Strukturänderungen durch LLMs und deren Metaeffekte Tätigkeiten in unbekannte kommunikative Umweltbedingungen katapultieren werden, die zusätzlich navigiert werden müssen.

Für Unternehmen eröffnen sich durch LLMs einerseits Möglichkeiten, Arbeitskosten zu sparen, indem man die Produktivität erhöht und dadurch entweder den Umsatz steigert oder Arbeitende freisetzt. Das kann bedeuten, dass z. B. Programmier*innen, Texter*innen, Grafiker*innen und Hotline-Angestellte ihre Arbeit verlieren.

Unternehmen können auch durch indirekten Einsatz von LLMs profitieren, etwa indem sie über KI die Urheberrechtsansprüche von Kreativen umgehen, wie es die Filmbranche in den USA gerade versucht (Broderick 2023). Es lohnt sich, genau hinzuschauen, welche Strategien hier zum Einsatz kommen: Oft geht es gar nicht um den Ersatz von Mitarbeiter*innen durch die KI selbst, die in vielen Belangen noch gar nicht fit genug für den Job ist, sondern die KI wird als Hebel genutzt – entweder um Angst unter den Mitarbeiter*innen zu schüren oder um Urheberrechtsansprüche

zu umgehen. Allgemein hilft ein angstgetriebener Diskurs zu KI der Arbeitgeberseite, den Wert menschlicher Arbeit infrage zu stellen.

Für Unternehmen ergeben sich eventuell auch neue Möglichkeiten der Kontrolle. LLMs könnten die Überwachung von unternehmensinterner Kommunikation vereinfachen. Semantisches Monitoring und automatisierte qualitative Beurteilung von Arbeitsergebnissen scheinen mit LLMs in Reichweite.

Gleichzeitig sind auch die Unternehmen selbst sehr viel angreifbarer durch Überwachung von LLMs. Insbesondere wenn sie LLMs als Dienst in der Cloud nutzen, wird eine Menge unternehmensinternen Wissens unwillkürlich an die Tech-Konzerne ausgelagert. Die geschaffenen Abhängigkeiten sind geeignet, die Unternehmen existenziell zu bedrohen.

Im größeren Zusammenhang: Wenn ein wachsender Teil der Wertschöpfung durch unternehmensexterne KIs der Tech-Konzerne geleistet wird, werden diese auch Wege finden, diesen Wert auch wieder abzuschöpfen. LLM einsetzende Unternehmen würden langfristig zu reinen Effektoren einer zentralisierten KI degradiert, die das Wissen tausender Unternehmen aufnimmt, verarbeitet und den Unternehmen dann wieder in optimierter Form als Service bereitstellt.

Und während sich Unternehmen am Anfang noch dagegen verwahren könnten, würde ihnen die zunehmend kompetitivere Marktsituation ab einem bestimmten Punkt keine Wahl mehr lassen, die effizienteren KI-Systeme zu implementieren – Systeme, die sie irgendwann nicht mal mehr infrage stellen können, weil sie die dafür nötigen Kompetenzen ja ausgelagert haben.

Ein zweiter Weg der LLMs in die Arbeitswelt führt durch die Hintertür und wird vorerst wahrscheinlich sogar den größten Teil der Adaption ausmachen. KI wird von Mitarbeiter*innen heimlich eingesetzt werden, das Management wird Mühe haben, den Einsatz überhaupt zu erkennen und zu regulieren. Aber das wird das Management tun müssen, wenn nicht Geschäftsgeheimnisse auf die Server von OpenAI geladen werden, ihre Mitarbeiter*innen Dokumente mit lauter Faktenfehlern verschicken (vgl. Weiser/Schweber 2023) oder die interne E-Mail-Kommunikation wegen exzessivem LLM-Einsatz zum Erliegen kommen soll.

Politik und Gewerkschaften werden den ersten Weg der LLMs adressieren können, den zweiten nur sehr schwer. Der heimliche Einsatz von LLMs wird populär sein und, weil der Chef/die Chefin davon sowieso nichts mitbekommt, können Mitarbeiter*innen die Produktivitätsgewinne zum Teil selbst einstreichen, sei es durch mehr Freizeit oder bessere Jobperformance. So zumindest erscheint es aus individuellen Perspektive.

Doch das ist ein Denkfehler. Da die Mitarbeiter*innen die LLMs im Kontext des Konkurrenzdrucks einsetzen, richtet sich der private Einsatz über

Bande auch gegen die eigenen Kolleg*innen. Am Ende sind alle mit erhöhten Produktivitätserwartungen konfrontiert.

Es scheint deswegen sinnvoll, neben den Versuchen der Regulierung und Eingrenzung des Einsatzes von LLMs auch eine Kampagne für einen offeneren Umgang mit diesen Werkzeugen in Betracht zu ziehen, inklusive einer Aufklärungskampagne über den sinnvollen und weniger sinnvollen Einsatz von LLMs. Allgemein wäre darüber nachzudenken, begleitende Hilfestellungen zum professionellen Umgang mit KI bereitzustellen. Das müsste in einem niedrigschwelligen und schnell aktualisierbaren Format geschehen, da die rasante Entwicklung in dem Bereich aktuelles Wissen schnell entwertet.

Für Gewerkschaften zeigt sich zudem eine weitere Herausforderung: Die betroffenen Berufsgruppen gehören nicht zum üblichen Klientel organisierter Interessenwahrnehmung. Aber das kann auch eine Chance sein, wenn es gelingt, diese Beschäftigten zu erreichen. Wenn es zu Protesten gegen den unternehmerischen KI-Einsatz kommt, könnten Gewerkschaften die Infrastruktur und das Know-how für diese Protestformen bereitstellen und langfristig für einen höheren Organisationsgrad in diesen Branchen sorgen.

Im öffentlichen Diskurs gilt es, den Apokalyptikern der KI-Revolution selbstbewusst entgegenzutreten. Menschliche Arbeit wird auch in Zukunft ihren Platz und ihren Wert behalten, Arbeitende haben ein Recht, die kommenden Strukturveränderungen mitgestalten zu dürfen. Es kann hilfreich sein, die Unternehmen daran zu erinnern, dass sie bei diesen Umwälzungen selbst leicht unter die Räder kommen könnten.

Ein Augenmerk für Gewerkschaften sollte auf den neuen Möglichkeiten der Mitarbeiterüberwachung durch LLMs liegen. Hier ergeben sich neuralgische Punkte, an denen man durch frühzeitige Gesetzgebung die Überwachungsmöglichkeiten in die Schranken weisen könnte.

Langfristig sollte man sich auf den Einzug und die Normalisierung dieser Werkzeuge und den sie begleitenden Strukturveränderungen und Metaeffekten gefasst machen. Diese werden nicht sofort sichtbar werden, sondern die Kommunikationserwartungen werden schleichend erodieren bzw. es werden sich neue ausbilden. Interessant ist dabei vor allem, sich wiederum auf die zu erwartenden Gegenbewegungen gefasst zu machen: die Wiederentdeckung des Telefonats, die neue Wertigkeit von Präsenzbegegnungen, die Abwertung von Fachjargon und generischem Text, die Nostalgie des Rechtschreibfehlers. Die Gegenbewegungen werden die Arbeitswelt ebenso stark prägen wie die LLMs selbst.

Es gibt noch so vieles, auf das das man sich gefasst machen muss. Deswegen wird das Fazit an dieser Stelle durch einen Epilog ergänzt.

7. Epilog

Die öffentliche Debatte um Künstliche Intelligenz geht sehr häufig um spekulative Szenarien rund um AGI, Superintelligenzen und die Frage, ob diese uns nun retten oder ausrotten werden. Doch LLMs müssen nicht superintelligent sein – eigentlich müssen sie überhaupt nicht in einem menschlichen Sinne intelligent sein –, um einen enormen Einfluss auf alle Aspekte unserer Welt zu haben.

Wenn eine Technologie so tief in unser kollektives Betriebssystem – die Sprache – implementiert wird, sind die Effekte vorhersehbar groß und unvorhersehbar vielfältig. Paul Virilio hat einmal gesagt, dass jede Technologie ihren eigenen Unfall produziert (Virilio/Lotringer 1983, S. 35 f.). Zwei Dinge sind dabei zu ergänzen: Ein Unfall ist nur dann ein Unfall, wenn er nicht vorhergesehen wurde. Und die Gefährlichkeit des Unfalls ist proportional zur Mächtigkeit des verunfallenden Systems. Denken wir an Social Media. Die Euphorie aus den Anfangstagen war im Nu verflogen, als wir feststellten, dass Plattformen als politische Waffen missbraucht werden können. Etwas sehr Ähnliches ist auch für LLMs zu erwarten.

Während diese Studie verfasst wurde, hat sich Elon Musk ausführlich über sein eigenes, geplantes LLM-Projekt namens xAI geäußert (Kerner 2023). Ob es jemals Realität wird, muss wie jede Ankündigung von Elon Musk in Zweifel gezogen werden (Molloy 2023). Dennoch lohnt es sich, die Rhetorik seiner Ankündigungen genauer zu betrachten. Musk spricht z. B. davon, dass sein LLM vor allem der „Wahrheit“ verpflichtet sein werde. Das hört sich erstmal gut an, denn die Wahrheit ist uns schließlich allen wichtig.

Nach allem, was wir in dieser Studie über LLMs gelernt haben, sollte uns diese Ankündigung aber auch sofort misstrauisch machen. LLMs sind strukturell nicht in der Lage, zwischen Wahrheit und Fiktion zu unterscheiden. Sie produzieren immer nur richtig ausschauende Antworten, die zwar häufig wahr sein können, aber nicht müssen. Auch wenn es wahrscheinlich Möglichkeiten gibt, dieses „Halluzinieren“ zu vermindern, ist der Anspruch eine „Wahrheits“-KI bauen zu wollen, ein enorm gefährlicher.

Schaut man weiter im Text, wird auch klar, was Musk genau meint. Seiner Ansicht nach müsse ein LLM von jeder „political correctness“ befreit werden, damit es fähig sei, auch „kontroverse Wahrheiten“ auszusprechen. Man muss das gar nicht in den Kontext seiner vielen rassistischen, sexistischen, trans- und homophoben und antisemitischen Äußerungen der letzten Jahre betrachten, um zu verstehen, was er damit sagen will. Es reicht, sich anzuschauen, wie er Twitter (mittlerweile „X“) führt und warum er es überhaupt gekauft hat. Twitter, die öffentlichste Bühne

unter den Internetplattformen, ist für Musk vor allem eine wichtigste Waffe im Kulturkampf geworden (Seemann 2023). xAI, wenn es je das Licht der Welt erblickt, soll einen sehr ähnlichen Zweck erfüllen.

Das Problem ist, dass sein Wunsch einer politisch unkorrekten KI sehr einfach zu erfüllen ist. Tatsächlich muss man viel Arbeit im Fine-Tuning aufwenden, um einem LLM zumindest die schlimmsten rassistischen Ausfälle halbwegs zuverlässig abzutrainieren. Spart man sich diese Arbeit, bekommt man sozusagen eine rassistische, sexistische und homophobe KI ab Werk.

Dafür gibt es ein bekanntes Beispiel. 2016 veröffentlichte Microsoft einen experimentellen Chatbot namens Tay, den es über die API mit Twitternutzer*innen interagieren ließ. Tay war so konfiguriert, dass es direkt aus den Konversationen mit anderen lernen konnte. Ein Teil der Nutzerschaft auf Twitter nutzte die Gelegenheit, Tay in allerlei Diskussionen über Rasse und Neonazismus zu verwickeln, bis Tay fast nur noch antisemitische, sexistische und rassistische Dinge ausspuckte (Vincent 2016).

Tay galt als eines der schlimmsten PR-Desaster in der jüngeren Microsoft-Geschichte und wurde als warnendes Beispiel verstanden. Elon Musk sieht darin wohl eher ein weiteres Beispiel der Cancel Culture und will mit xAI diese Leerstelle wieder füllen. Wenn seine KI dereinst schwarze Menschen beleidigt oder von der jüdischen Weltverschwörung redet, wird er das nicht als Fehler betrachten, sondern als „die Wahrheit" deklarieren. Seine große Gefolgschaft an jungen, weißen Männern wird auch diesmal applaudieren.

Noch einmal: Es ist nicht abzusehen, ob das Projekt überhaupt veröffentlicht wird oder ob es so kommen wird wie oben beschrieben. Doch Musks Plan weist auf eine Gefahr hin, die noch zu wenig thematisiert wird: Es ist nicht nur so, dass LLMs die Biases der Menschen übernehmen oder problematische Denkfiguren reproduzieren. Manche Menschen könnten das genauso wollen.

Dass mit Sprache Politik gemacht wird, ist keine Neuigkeit. Jede Verwendung von Sprache ist zumindest auch politisch, reproduziert sie doch unwillkürlich all die Muster, Narrative und Figuren, auf die wir in der Kommunikation unbewusst zurückgreifen. So definiert z. B. jeder Sprachakt immer auch mit, wo die Grenze zwischen dem verläuft, was eine normale, legitime Äußerung ist, und was nicht (vgl. Mackinac Center for Public Policy 2023).

Ein LLM, zumindest wenn es von vielen Menschen im Alltag verwendet wird, ist eine Teilautomatisierung von Aussagen. Es produziert Sprachakte am Fließband, die von Menschen oft ohne viel Reflexion übernommen und weiterverbreitet werden. LLMs können darüber hinaus ganz automatisiert die Kommunikationswege befüllen und tun das bereits. Wenn

jemand Interesse daran hat, eine bestimmte Sprachfigur zu etablieren oder eine bestimmte Rhetorik zu normalisieren, dann wäre die Kontrolle über ein populäres LLM enorm praktisch.

So könnte eine mögliche Zukunft von LLMs aussehen: Politisch segregiert nutzen wir das eine, aber nicht das andere LLM, nicht nur um unsere Kommunikation und unsere Arbeits- und Denkprozesse zu beschleunigen, sondern auch um unsere Sicht auf die Welt auszudrücken. Wenden Sie sich deswegen einfach an den einen oder an den anderen Tech-Konzern ihres Vertrauens.

Schaut man sich Phänomene wie QAnon an, ist es sogar leicht vorstellbar, dass sich um bestimmte LLMs ganze politische Bewegungen, vielleicht sogar sektenartige Anhänger*innen versammeln, die in dem Output der Maschine die Offenbarung einer höheren spirituellen Wahrheit wähnen. Dafür müssten sich LLMs technisch gar nicht weiter entwickeln. Im Gegenteil, ein zu kohärenter Output wäre hier sowieso nur hinderlich.

Oder es könnte ganz anders kommen, und es formiert sich eine gesellschaftliche Gegenmacht, die grundlegende Neuausrichtung unserer gesellschaftlichen Kommunikationsstruktur durch einige wenige Internetkonzerne nicht hinzunehmen bereit ist. Es könnte sich ein breiter Widerstand formen, der versucht, über öffentliche Proteste und politische Einflussnahme die Weiterentwicklung von solchen oder ähnlichen KI-Systemen zu stoppen. Prominente, Politiker*innen und Institutionen könnten sich selbst verpflichten, diese Systeme zu boykottieren. Es könnten Verbote von KI in bestimmten Bereichen der menschlichen Kommunikation erlassen werden (Geuter 2023); es könnte vielleicht die Technologie selbst verboten und international geächtet werden (Reijers/Maschewski/Nosthoff 2023).

Auch dieses Szenario ist absolut vorstellbar, wenn man bedenkt, dass im Vergleich zu den bisherigen Automatisierungswellen diese Welle eine wirtschaftlich gut aufgestellte, medial kompetente und sozial gut vernetzte Gruppe bedroht.

Auch über die Frage der Arbeitswelt hinaus ist das Thema Large Language Model ein politisches Thema. Es ist ein Missstand, dass es nach wie vor unter vornehmlich technischen Gesichtspunkten verhandelt wird.

Literatur

The Ad Hoc Committee on the Triple Revolution (1964): The Triple Revolution: Cybernation, Weaponry, Human Rights. www.educationanddemocracy.org/FSCfiles/C_CC2a_TripleRevolution.htm (Abruf am 18.8.2023).

Albrecht, Steffen (2023): ChatGPT und andere Computermodelle zur Sprachverarbeitung – Grundlagen, Anwendungspotenziale und mögliche Auswirkungen. Büro für Technikfolgen-Abschätzung beim Deutschen Bundestag. www.bundestag.de/resource/blob/944148/30b0896f6e49908155fcd01d77f57922/20-18-109-Hintergrundpapier-data.pdf (Abruf am 18.8.2023).

Ali, Rohaid / Tang, Oliver Y. / Connolly, Ian D. / Fridley, Jared S. / Shin, John H. / Zadnik Sullivan, Patricia L. / Cielo, Deus / Oyelese, Adetokunbo A. / Doberstein, Curtis E. / Telfeian, Albert E. / Gokaslan, Ziya L. / Asaad, Wael F. (2023): Performance of ChatGPT, GPT-4, and Google Bard on a Neurosurgery Oral Boards Preparation Question Bank. In: medRxiv, 12.4.2023. DOI: 10.1101/2023.04.06.23288265 (Abruf am 18.8.2023).

Alloway, Tracy / Weisenthal, Joe / Ruffin, Abraiya (2023): An Early Investor in AI Explains How They're Thinking About the Space Now. In: Bloomberg, 17.7.2023. www.bloomberg.com/news/articles/2023-07-17/an-early-investor-in-ai-explains-how-they-re-thinking-about-the-space-now?srnd=oddlots#xj4y7vzkg (Abruf am 18.8.2023).

Andreenko, Artem (2023): Tweet vom 12.3.2023. https://twitter.com/miolini/status/1634982361757790209 (Abruf am 18.8.2023).

Andreessen, Marc (2023): Why AI Will Save the World. https://a16z.com/2023/06/06/ai-will-save-the-world/ (Abruf am 18.8.2023).

Anthropic (2023): Introducing Claude. www.anthropic.com/index/introducing-claude (Abruf am 18.8.2023).

Arnold, Chris (2017): Tax Bill Favors Adding Robots Over Workers, Critics Say. In: National Public Radio, 8.12.2017. www.npr.org/2017/12/08/569118310/tax-bill-favors-adding-robots-over-workers-critics-say (Abruf am 18.8.2023).

Arntz, Melanie / Gregory, Terry / Zierahn, Ulrich (2023): The Risk of Automation for Jobs in OECD Countries. A Comparative Analysis, Paris. www.oecd-ilibrary.org/social-issues-migration-health/the-risk-of-automation-for-jobs-in-oecd-countries_5jlz9h56dvq7-en (Abruf am 18.8.2023).

Autor, David H. (2015): Why Are There Still So Many Jobs? The History and Future of Workplace Automation. In: Journal of Economic Perspectives 29, H. 3, S. 3–30. DOI: 10.1257/jep. 29.3.3 (Abruf am 18.8.2023).

Bajohr, Hannes (2023a): Dumb Meaning: Machine Learning and Artificial Semantics. In: Image. The Interdisciplinary Journal of Image Sciences 37, H. 1, S. 58–70.

Bajohr, Hannes (2023b): On Artificial and Post-Artificial Texts. https://hannesbajohr.de/en/2023/03/11/on-artificial-and-post-artificial-texts/ (Abruf am 18.8.2023).

Banner, Ron / Nahshan, Yury / Hoffer, Elad / Soudry, Daniel (2018): Post-training 4-bit quantization of convolution networks for rapid-deployment. https://arxiv.org/pdf/1810.05723 (Abruf am 18.8.2023).

Barr, Alistair (2023): The world's most powerful AI model suddenly got "lazier" and "dumber." A radical redesign of OpenAI's GPT-4 could be behind the decline in performance. In: Business Insider, 12.7.2023. www.businessinsider.com/openai-gpt4-ai-model-got-lazier-dumber-chatgpt-2023-7 (Abruf am 18.8.2023).

Barthes, Roland (2000): Der Tod des Autors. In: Jannidis, Fotis / Lauer, Gerhard / Martínez, Matías / Winko, Simone (Hrsg.): Texte zur Theorie der Autorschaft, Ditzingen: Reclam, S. 185–193.

Beckett, Lois (2023): "Those who hate AI are insecure": inside Hollywood's battle over artificial intelligence. In: The Guardian, 26.5.2023. www.theguardian.com/us-news/2023/may/26/hollywood-writers-strike-artificial-intelligence (Abruf am 18.8.2023).

Belkin, Mikhail / Hsu, Daniel / Ma, Siyuan / Mandal, Soumik (2019): Reconciling modern machine-learning practice and the classical bias-variance trade-off. In: Proceedings of the National Academy of Sciences of the United States of America 116, H. 32, S. 15849–15854.

Benanav, Aaron (2023): The revolution will not be brought to you by ChatGPT. In: The New Statesman, 11.4.2023. www.newstatesman.com/ideas/2023/04/revolution-brought-chatgpt-artificial-intelligence (Abruf am 18.8.2023).

Bender, Emily M. / Koller, Alexander (2020): Climbing towards NLU. On Meaning, Form, and Understanding in the Age of Data. In: Jurafsky, Dan / Chai, Joyce / Schluter, Natalie / Tetreault, Joel (Hrsg.): Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Stroudsburg, Pa.: Association for Computational Linguistics, S. 5185–5198.

Bender, Emily M. / Gebru, Timnit / McMillan-Major, Angelina / Shmitchell, Shmargaret (2021): On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, New York: Association for Computing Machinery, S. 610–623. DOI: 10.1145/3442188.3445922 (Abruf am 18.8.2023).

Bogost, Ian (2022): ChatGPT Is Dumber Than You Think. Treat it like a toy, not a tool. In: The Atlantic, 7.12.2022. www.theatlantic.com/technology/archive/2022/12/chatgpt-openai-artificial-intelligence-writing-ethics/672386/ (Abruf am 18.8.2023).

Boussioux, Leonard / Lane, Jacqueline N. / Zhang, Miaomiao / Jacimovic, Vladimir / Lakhani, Karim R. (2023) The Crowdless Future? How Generative AI Is Shaping the Future of Human Crowdsourcing. Harvard Business School Technology & Operations Mgt. Unit Working Paper No. 24–005, 7.8.203. https://ssrn.com/abstract=4533642 (Abruf am 18.8.2023).

Bove, Tristan (2023): CEO of Google's DeepMind says we could be "just a few years" from A. I. that has human-level intelligence. In: Fortune, 3.5.2023. https://fortune.com/2023/05/03/google-deepmind-ceo-agi-artificial-intelligence/ (Abruf am 18.8.2023).

Box, George E. P. (1979): Robustness in the strategy of scientific model building. In: Launer, Robert L. / Wilkinson, Graham N. (Hrsg.): Robustness in Statistics, Cambridge, Mass.: Academic Press, S. 201–236. https://doi.org/10.1016/b978-0-12-438150-6.50018-2 (Abruf am 4.9.2023).

Broderick, Ryan (2023): AI can't replace humans yet – but if the WGA writers don't win, it might not matter. In: Polygon, 31.5.2023. www.polygon.com/23742770/ai-writers-strike-chat-gpt-explained (Abruf am 18.8.2023).

Brunner, Katharina / Harlan, Elisa (2023): We Are All Raw Material for AI. In: Bayerischer Rundfunk, 7.7.2023. https://interaktiv.br.de/ki-trainingsdaten/en/index.html (Abruf am 18.8.2023).

Brynjolfsson, Erik / McAfee, Andrew (2014): The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies. Narrated by Jeff Cummings, Grand Haven, Mich.: Brilliance Audio.

Bubeck, Sébastien / Chandrasekaran, Varun / Eldan, Ronen / Gehrke, Johannes / Horvitz, Eric / Kamar, Ece / Lee, Peter / Lee, Yin Tat / Li, Yuanzhi / Lundberg, Scott / Nori, Harsha / Palangi, Hamid / Ribeiro, Marco Tulio / Zhang, Yi (2023): Sparks of Artificial General Intelligence: Early experiments with GPT-4. https://arxiv.org/pdf/2303.12712 (Abruf am 18.8.2023).

Cai, Kenrick / Martin, Iain (2023): The AI Founder Taking Credit For Stable Diffusion's Success Has A History Of Exaggeration. In: Forbes, 4.6.2023. www.forbes.com/sites/kenrickcai/2023/06/04/stable-diffusion-emad-mostaque-stability-ai-exaggeration/?sh=3f20e01a75c5 (Abruf am 18.8.2023).

Carstensen, Tanja / Ganz, Kathrin (2023): Vom Algorithmus diskriminiert? Zur Aushandlung von Gender in Diskursen über Künstliche Intelligenz und Arbeit, Working Paper 274, Düsseldorf: Hans-Böckler-Stiftung. www.boeckler.de/de/faust-detail.htm?sync_id=HBS-008607 (Abruf am 18.8.2023).

Che, Chang / Wang, Olivia (2023): What Happens When You Ask a Chinese Chatbot About Taiwan? In: New York Times, 14.6.2023. www.nytimes.com/2023/07/14/business/baidu-ernie-openai-chatgpt-chinese.html (Abruf am 18.8.2023).

Chiang, Ted (2023): ChatGPT Is a Blurry JPEG of the Web. In: The New Yorker, 9.2.2023. www.newyorker.com/tech/annals-of-technology/chatgpt-is-a-blurry-jpeg-of-the-web (Abruf am 18.8.2023).

Chomsky, Noam / Roberts, Ian / Watumull, Jeffrey (2023): Noam Chomsky: The False Promise of ChatGPT. In: Portside, 3.8.2023. https://portside.org/2023-03-08/noam-chomsky-false-promise-chatgpt (Abruf am 18.8.2023).

Claburn, Thomas (2023): How prompt injection attacks hijack today's top-end AI – and it's tough to fix. In: The Register, 26.4.2023. www.theregister.com/2023/04/26/simon_willison_prompt_injection/ (Abruf am 18.8.2023).

Creative Destruction Lab (2016): Geoff Hinton: On Radiology. In: Youtube, 24.11.2016. www.youtube.com/watch?v=2HMPRXstSvQ (Abruf am 18.8.2023).

Davenport, Thomas H. / Miller, Steven M. (2022): Working with AI. Real stories of human-machine collaboration, Cambridge, Mass. / London: The MIT Press.

Davis, Wes (2023): Sarah Silverman is suing OpenAI and Meta for copyright infringement. In: The Verge, 9.7.2023. www.theverge.com/2023/7/9/23788741/sarah-silverman-openai-meta-chatgpt-llama-copyright-infringement-chatbots-artificial-intelligence-ai (Abruf am 18.8.2023).

Derrida, Jacques (1983): Grammatologie, Frankfurt am Main: Suhrkamp.

Derrida, Jacques (1988): Die Différance. In: Derrida, Jacques (Hrsg.): Randgänge der Philosophie, Wien: Passagen, S. 31–56.

Dertat, Arden (2017): Applied Deep Learning – Part 1: Artificial Neural Networks. In: Towards Data Science, 8.8.2017. https://towardsdatascience.com/applied-deep-learning-part-1-artificial-neural-networks-d7834f67a4f6 (Abruf am 18.8.2023).

Di Battista, Attilio / Grayling, Sam / Hasselaar, Elselot / Leopold, Till / Rayner, Mark / Zahidi, Saadia (2023): Future of Jobs Report 2023. Insight Report, Geneva: World Economic Forum. www.weforum.org/reports/the-future-of-jobs-report-2023 (Abruf am 18.8.2023).

Diering, Frank (2023): Der Mann, der ChatGPT erfand. In: Die Welt, 15.3.2023. www.welt.de/kultur/article244204855/Der-Mann-der-ChatGPT-erfand.html (Abruf am 18.8.2023).

Doshi, Anil Rajnikant / Hauser, Oliver (2023): Generative artificial intelligence enhances creativity. https://ssrn.com/abstract=4535536 (Abruf am 18.8.2023).

Dudley, Simon (2014): The Internet Just Isn't That Big a Deal Yet. A Hard Look at Solow's Paradox. In: Wired. www.wired.com/insights/2014/11/solows-paradox/ (Abruf am 18.8.2023).

Dzieza, Josh (2023): AI Is a Lot of Work. In: The Verge, 20.6.2023. www.theverge.com/features/23764584/ai-artificial-intelligence-data-notation-labor-scale-surge-remotasks-openai-chatbots (Abruf am 18.8.2023).

Eloundou, Tyna / Manning, Sam / Mishkin, Pamela / Rock, Daniel (2023): GPTs are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models. https://arxiv.org/pdf/2303.10130 (Abruf am 18.8.2023).

European Parliament (2023): EU AI Act: first regulation on artificial intelligence. In: European Parliament News. www.europarl.europa.eu/news/en/headlines/society/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence (Abruf am 18.8.2023).

Felten, Edward W. / Raj, Manav / Seamans, Robert (2023): Occupational Heterogeneity in Exposure to Generative AI. In: SSRN Electronic Journal. DOI: 10.2139/ssrn.4414065 (Abruf am 18.8.2023).

Frenkel, Sheera / Thompson, Stuart A. (2023): "Not for Machines to Harvest": Data Revolts Break Out Against A. I. In: The New York Times, 15.7.2023. www.nytimes.com/2023/07/15/technology/artificial-intelligence-models-chat-data.html (Abruf am 18.8.2023).

Frey, Carl Benedikt / Osborne, Michael A. (2017): The future of employment: How susceptible are jobs to computerisation? In: Technological Forecasting and Social Change 114, S. 254–280. DOI: 10.1016/j.techfore.2016.08.019 (Abruf am 18.8.2023).

Future of Life Institute (2015): Research Priorities for Robust and Beneficial Artificial Intelligence: An Open Letter. https://futureoflife.org/open-letter/ai-open-letter/ (Abruf am 18.8.2023).

Future of Life Institute (2023): Pause Giant AI Experiments: An Open Letter. https://futureoflife.org/open-letter/pause-giant-ai-experiments/ (Abruf am 18.8.2023).

Futurezone (2023): Microsoft stellt Twitter-Werbung ein, Musk droht mit Klage. In: Futurezone, 20.4.2023. https://futurezone.at/b2b/microsoft-smart-campaign-twitter-werbung-api-schnittstelle-elon-musk-klage-openai/402418751 (Abruf am 18.8.2023).

Geuter, Jürgen (2022): Das Dritte Web. https://tante.cc/2022/02/04/das-dritte-web/ (Abruf am 18.8.2023).

Geuter, Jürgen (2023): „I'm sorry HAL, I won't let you do that." Vortrag auf der re:publica 2023. https://tante.cc/2023/06/06/im-sorry-hal-i-wont-let-you-do-that-vortrag-auf-der-republica-2023/ (Abruf am 18.8.2023).

Ghahramani, Zoubin (2023): Introducing PaLM 2. https://blog.google/technology/ai/google-palm-2-ai-large-language-model/ (Abruf am 18.8.2023).

Gilardi, Fabrizio / Alizadeh, Meysam / Kubli, Maël (2023): ChatGPT outperforms crowd workers for text-annotation tasks. https://arxiv.org/pdf/2303.15056 (Abruf am 18.8.2023).

Gimpel, Henner / Hall, Kristina / Decker, Stefan / Eymann, Torsten / Lämmermann, Luis / Mädche, Alexander / Röglinger, Maximilian / Ruiner, Caroline / Schoch, Manfred / Schoop, Mareike / Urbach, Nils / Vandirk, Steffen (2023): Unlocking the Power of Generative AI Models and Systems such as GPT-4 and ChatGPT for Higher Education. A Guide for Students and Lecturers, Stuttgart. https://digital.uni-hohenheim.de/fileadmin/einrichtungen/digital/Generative_AI_and_ChatGPT_in_Higher_Education.pdf (Abruf am 18.8.2023).

Girotra, Karan / Meincke, Lennart / Terwiesch, Christian / Ulrich, Karl T. (2023): Ideas are Dimes a Dozen. Large Language Models for Idea Generation in Innovation. https://ssrn.com/abstract=4526071 (Abruf am 18.8.2023).

Goddard, Kate / Roudsari, Abdul / Wyatt, Jeremy C. (2012): Automation bias: a systematic review of frequency, effect mediators, and mitigators. In: Journal of the American Medical Informatics Association 19, H. 1, S. 121–127. https://pubmed.ncbi.nlm.nih.gov/21685142/ (Abruf am 18.8.2023).

Goldman, Sharon (2023): AI experts challenge “doomer” narrative, including “extinction risk” claims. In: VentureBeat, 31.5.2023. https://venturebeat.com/ai/ai-experts-challenge-doomer-narrative-including-extinction-risk-claims/ (Abruf am 18.8.2023).

Google (2023): PaLM 2 Technical Report. https://arxiv.org/abs/2305.10403 (Abruf am 18.8.2023)

Gordon, Robert J. (2016): The rise and fall of American growth. The US standard of living since the Civil War, Princeton / Oxford: Princeton University Press.

Grant, Nico / Metz, Cade (2022): A New Chat Bot Is a “Code Red” for Google’s Search Business. In: The New York Times, 21.12.2022. www.nytimes.com/2022/12/21/technology/ai-chatgpt-google-search.html (Abruf am 18.8.2023).

Gudibande, Arnav / Wallace, Eric / Snell, Charlie / Geng, Xinyang / Liu, Hao / Abbeel, Pieter / Levine, Sergey / Song, Dawn (2023): The False Promise of Imitating Proprietary LLMs. https://arxiv.org/abs/2305.15717 (Abruf am 18.8.2023).

Hahn, Silke (2023): Claude 2 verarbeitet ganze Bücher in Sekunden. In: Spektrum der Wissenschaft, 13.7.2023. www.spektrum.de/news/anthropic-baut-mit-ki-system-claude-2-vorsprung-in-der-textanalyse-aus/2158890 (Abruf am 18.8.2023).

Hanna, Alex / Bender, Emily M. (2023): “AI” Hurts Consumers and Workers – and Isn’t Intelligent. In: Tech Policy Press, 4.8.2023. https://techpolicy.press/ai-hurts-consumers-and-workers-and-isnt-intelligent/ (Abruf am 18.8.2023).

Harnad, Stevan (1990): The Symbol Grounding Problem. In: Physica D 42, S. 335–346. https://web-archive.southampton.ac.uk/cogprints.org/3106/01/sgproblem1.html (Abruf am 18.8.2023).

Hatzius, Jan / Briggs, Joseph / Kodnani, Devesh / Pierdomenico, Giovanni (2023): The Potentially Large Effects of Artificial Intelligence on Economic Growth. Goldman Sachs. www.gspublishing.com/content/research/en/reports/2023/03/27/d64e052b-0f6e-45d7-967b-d7be35fabd16.html (Abruf am 18.8.2023).

Heaven, Will Douglas (2023): Geoffrey Hinton tells us why he's now scared of the tech he helped build. In: MIT Technology Review, 2.5.2023. www.technologyreview.com/2023/05/02/1072528/geoffrey-hinton-google-why-scared-ai/ (Abruf am 18.8.2023).

Herbold, Steffen / Hautli-Janisz, Annette / Heuer, Ute / Kikteva, Zlata / Trautsch, Alexander (2023): AI, write an essay for me: A large-scale comparison of human-written versus ChatGPT-generated essays. https://arxiv.org/pdf/2304.14276 (Abruf am 18.8.2023).

Hoffmann, Jordan / Borgeaud, Sebastian / Mensch, Arthur / Buchatskaya, Elena / Cai, Trevor / Rutherford, Eliza / Casas, Diego de Las / Hendricks, Lisa Anne / Welbl, Johannes / Clark, Aidan / Hennigan, Tom / Noland, Eric / Millican, Katie / van den Driessche, George / Damoc, Bogdan / Guy, Aurelia / Osindero, Simon / Simonyan, Karen / Elsen, Erich / Rae, Jack W. / Vinyals, Oriol / Sifre, Laurent (2022): Training Compute-Optimal Large Language Models. https://arxiv.org/pdf/2203.15556 (Abruf am 18.8.2023).

Huggingface Leaderboard (2023): Open LLM Leaderboard. https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard (Abruf am 18.8.2023).

ILO/OECD (2015): The Labour Share in G20 Economies. International Labour Organization/Organisation for Economic Co-operation and Development. www.oecd.org/g20/topics/employment-and-social-policy/The-Labour-Share-in-G20-Economies.pdf (Abruf am 18.8.2023).

Jiao, Wenxiang / Wang, Wenxuan / Huang, Jen-tse / Wang, Xing / Tu, Zhaopeng (2023): Is ChatGPT A Good Translator? Yes With GPT-4 As The Engine. https://arxiv.org/abs/2301.08745 (Abruf am 18.8.2023).

Jin, Charles / Rinard, Martin (2023): Evidence of Meaning in Language Models Trained on Programs. https://arxiv.org/pdf/2305.11169 (Abruf am 18.8.2023).

Jumper, John / Evans, Richard / Pritzel, Alexander / Green, Tim / Figurnov, Michael / Ronneberger, Olaf / Tunyasuvunakool, Kathryn / Bates, Russ/Žídek, Augustin / Potapenko, Anna / Bridgland, Alex / Meyer, Clemens / Kohl, Simon A. A. / Ballard, Andrew J. / Cowie, Andrew / Romera-Paredes, Bernardino / Nikolov, Stanislav / Jain, Rishub / Adler, Jonas / Back, Trevor / Petersen, Stig / Reiman, David / Clancy, Ellen / Zielinski, Michal / Steinegger, Martin / Pacholska, Michalina / Berghammer, Tamas / Bodenstein, Sebastian / Silver, David / Vinyals, Oriol / Senior, Andrew W. / Kavukcuoglu, Koray / Kohli, Pushmeet / Hassabis, Demis (2021): Highly accurate protein structure prediction with AlphaFold. In: Nature 596, S. 583–589. DOI: 10.1038/s41586-021-03819-2 (Abruf am 18.8.2023).

Kalliamvakou, Eirini (2022): Research: quantifying GitHub Copilot's impact on developer productivity and happiness. https://github.blog/2022-09-07-research-quantifying-github-copilots-impact-on-developer-productivity-and-happiness/ (Abruf am 18.8.2023).

Kaplan, Jared / McCandlish, Sam / Henighan, Tom / Brown, Tom B. / Chess, Benjamin / Child, Rewon / Gray, Scott / Radford, Alec / Wu, Jeffrey / Amodei, Dario (2020): Scaling Laws for Neural Language Models. https://arxiv.org/abs/2001.08361v1 (Abruf am 18.8.2023).

Karpathy, Andrej (2023): State of GPT. Microsoft Build. https://build.microsoft.com/en-US/sessions/db3f4859-cd30-4445-a0cd-553c3304f8e2 (Abruf am 18.8.2023).

Katz, Daniel Martin / Bommarito, Michael James / Gao, Shang / Arredondo, Pablo (2023): GPT-4 Passes the Bar Exam. In: SSRN Electronic Journal. DOI: 10.2139/ssrn.4389233 (Abruf am 18.8.2023).

Kerner, Sean Michael (2023): Elon Musk reveals xAI efforts, predicts full AGI by 2029. In: VentureBeat, 15.7.2023. https://venturebeat.com/ai/elon-musk-reveals-xai-efforts-predicts-full-agi-by-2029/ (Abruf am 18.8.2023).

Keynes, John Maynard (1932): Economic Possibilities for our Grandchildren (1930). In: Keynes, John Maynard (Hrsg.): Essays in Persuasion, New York: Harcourt Brace, S. 358–373. www.aspeninstitute.org/wp-content/uploads/files/content/upload/Intro_and_Section_I.pdf (Abruf am 18.8.2023).

Klein, Ezra (2023): Beyond the "Matrix" Theory of the Human Mind. In: The New York Times, 28.5.2023. www.nytimes.com/2023/05/28/opinion/artificial-intelligence-thinking-minds-concentration.html (Abruf am 18.8.2023).

Kleist, Heinrich von (1805): Über die allmähliche Verfertigung der Gedanken beim Reden. Projekt Gutenberg. www.projekt-gutenberg.org/kleist/gedanken/gedanken.html (Abruf am 18.8.2023).

Knight, Will (2023): A New Attack Impacts ChatGPT – and No One Knows How to Stop It. In: Wired, 1.8.2023. www.wired.com/story/ai-adversarial-attacks/ (Abruf am 18.8.2023).

Kramer, Josefine (2023): EU will Auskunft über die Trainingsdaten von ChatGPT. In: t3n Magazin, 28.4.2023. https://t3n.de/news/ai-act-eu-trainingsdaten-chatgpt-urheberrecht-1549442/ (Abruf am 18.8.2023).

Kurzweil, Ray (2005): The singularity is near. When humans transcend biology, New York, N. Y.: Penguin Books.

Larsen, Benjamin Cedric (2022): The geopolitics of AI and the rise of digital sovereignty. In: Brookings Institute, 8.12.2022. www.brookings.edu/articles/the-geopolitics-of-ai-and-the-rise-of-digital-sovereignty/ (Abruf am 18.8.2023).

Lee, Timothy B. / Trott, Sean (2023): Large language models, explained with a minimum of math and jargon. In: Understanding AI, 27.7.2023. www.understandingai.org/p/large-language-models-explained-with (Abruf am 18.8.2023).

Li, Kenneth / Hopkins, Aspen K. / Bau, David / Viégas, Fernanda / Pfister, Hanspeter / Wattenberg, Martin (2023): Emergent World Representations: Exploring a Sequence Model Trained on a Synthetic Task. https://arxiv.org/abs/2210.13382v4 (Abruf am 18.8.2023).

Liang, Weixin / Yuksekgonul, Mert / Mao, Yining / Wu, Eric / Zou, James (2023): GPT detectors are biased against non-native English writers. In: Patterns 4, H. 7, S. 100779. DOI: 10.1016/j.patter.2023.100779 (Abruf am 18.8.2023).

LMSYS Org (2023a): Chatbot Arena. https://chat.lmsys.org/?arena (Abruf am 18.8.2023).

LMSYS Org (2023b): Leaderboard. https://huggingface.co/spaces/lmsys/chatbot-arena-leaderboard (Abruf am 18.8.2023).

Loiseau, Jean-Christophe B. (2019): Rosenblatt's perceptron, the first modern neural network. A quick introduction to deep learning for beginners. In: Towards Data Science, 11.3.2019. https://towardsdatascience.com/rosenblatts-perceptron-the-very-first-neural-network-37a3ec09038a (Abruf am 18.8.2023).

Luitse, Dieuwertje / Denkena, Wiebke (2021): The great Transformer: Examining the role of large language models in the political economy of AI. In: Big Data & Society 8, H. 2, S. 1–14. DOI: 10.1177/20539517211047734 (Abruf am 18.8.2023).

Mackinac Center for Public Policy (2023): The Overtone Window. www.mackinac.org/OvertonWindow (Abruf am 18.8.2023).

Maddaus, Gene (2023): WGA Would Allow Artificial Intelligence in Scriptwriting, as Long as Writers Maintain Credit. In: Variety, 22.3.2023. https://variety.com/2023/biz/news/writers-guild-artificial-intelligence-proposal-1235560927/ (Abruf am 18.8.2023).

Mahowald, Kyle / Ivanova, Anna A. / Blank, Idan A. / Kanwisher, Nancy / Tenenbaum, Joshua B. / Fedorenko, Evelina (2023): Dissociating language and thought in large language models: a cognitive perspective. https://arxiv.org/pdf/2301.06627 (Abruf am 18.8.2023).

Manyika, James / Lund, Susan / Chui Michael/Bughin Jacques/Woetzel, Jonathan / Batra, Parul / Ko, Ryan / Sanghvi, Saurabh (2017): Jobs lost, jobs gained: What the future of work will mean for jobs, skills, and wages. McKinsey. www.mckinsey.com/featured-insights/future-of-work/jobs-lost-jobs-gained-what-the-future-of-work-will-mean-for-jobs-skills-and-wages (Abruf am 18.8.2023).

Marcus, Gary (2022): Deep Learning Is Hitting a Wall. In: Nautilus, 10.3.2022. https://nautil.us/deep-learning-is-hitting-a-wall-238440/ (Abruf am 18.8.2023).

Marcus, Gary (2023): The Sparks of AGI? Or the End of Science? Marching into the future with an obstructed view. In: Marcus on AI, 24.3.2023. https://garymarcus.substack.com/p/the-sparks-of-agi-or-the-end-of-science (Abruf am 18.8.2023).

Marx, Karl (1867/1980): Das Kapital. Band 1, Berlin: Dietz.

Matt (2018): Can a Robot Write a Symphony? In: Know Your Meme. https://knowyourmeme.com/memes/can-a-robot-write-a-symphony (Abruf am 18.8.2023).

Mayo, Benjamin (2023): Apple says it has fixed iPhone autocorrect with iOS 17. In: 9TO5Mac, 5.6.2023. https://9to5mac.com/2023/06/05/ios-17-iphone-autocorrect/ (Abruf am 18.8.2023).

McCulloch, Warren S. / Pitts Walter (1943/1990): Logical Calculus of the Ideas Immanent in Nervous Activity. In: Bulletin of Mathematical Biology 52, H. I/2, S. 99–115. www.cs.cmu.edu/~./epxing/Class/10715/reading/McCulloch.and.Pitts.pdf (Abruf am 18.8.2023).

McKenzie, Ian R. / Lyzhov, Alexander / Pieler, Michael / Parrish, Alicia / Mueller, Aaron / Prabhu, Ameya / McLean, Euan / Kirtland, Aaron / Ross, Alexis / Liu, Alisa / Gritsevskiy, Andrew / Wurgaft, Daniel / Kauffman, Derik / Recchia, Gabriel / Liu, Jiacheng / Cavanagh, Joe / Weiss, Max / Huang, Sicong / Droid, The Floating / Tseng, Tom / Korbak, Tomasz / Shen, Xudong / Zhang, Yuhui / Zhou, Zhengping / Kim, Najoung / Bowman, Samuel R. / Perez, Ethan (2023): Inverse Scaling: When Bigger Isn't Better. https://arxiv.org/abs/2306.09479 (Abruf am 18.8.2023).

Meta (2023a): Introducing LLaMA: A foundational, 65-billion-parameter large language model. In: Meta AI Blog. https://ai.facebook.com/blog/large-language-model-llama-meta-ai/ (Abruf am 18.8.2023).

Meta (2023b): Meta and Microsoft Introduce the Next Generation of Llama. In: Meta Newsroom. https://about.fb.com/news/2023/07/llama-2/ (Abruf am 18.8.2023).

Mikolov, Tomas / Sutskever, Ilya / Chen, Kai / Corrado, Greg / Dean, Jeffrey (2013): Distributed Representations of Words and Phrases and their Compositionality. In: Advances in Neural Information Processing Systems 26, S. 3111–3119.

Mollick, Ethan (2023a): Detecting the Secret Cyborgs. In: One Useful Thing. www.oneusefulthing.org/p/detecting-the-secret-cyborgs (Abruf am 18.8.2023).

Mollick, Ethan (2023b): Setting time on fire and the temptation of The Button. In: One Useful Thing. www.oneusefulthing.org/p/setting-time-on-fire-and-the-temptation (Abruf am 18.8.2023).

Mollick, Ethan (2023c): The Homework Apocalypse. In: One Useful Thing. www.oneusefulthing.org/p/the-homework-apocalypse (Abruf am 18.8.2023).

Molloy, Parker (2023): Vaporware King Elon Musk's xAI is Basically Just a 2023 Version of Microsoft's Tay. In: The Present Age. www.readtpa.com/p/vaporware-king-elon-musks-xai-is (Abruf am 18.8.2023).

Morozov, Evgeny (2023): The True Threat of Artificial Intelligence. In: The New York Times, 30.6.2023. www.nytimes.com/2023/06/30/opinion/artificial-intelligence-danger.html (Abruf am 18.8.2023).

Nagyfi, Richard (2018): The differences between Artificial and Biological Neural Networks. In: Towards Data Science, 4.9.2018. https://towardsdatascience.com/the-differences-between-artificial-and-biological-neural-networks-a8b46db828b7 (Abruf am 18.8.2023).

Narayanan, Arvind / Kapoor, Sayash (2023): GPT-4 and professional benchmarks: the wrong answer to the wrong question. In: AI Snake Oil, 20.3.2023. https://aisnakeoil.substack.com/p/gpt-4-and-professional-benchmarks (Abruf am 18.8.2023).

Natale, Simone (2021): Deceitful Media. Artificial Intelligence and Social Life after the Turing Test, Oxford: Oxford University Press.

Neumann, Uwe (2023): Regional adaptability to digital change. May the Swabian force be with you, Essen: RWI – Leibniz-Institut für Wirtschaftsforschung. www.rwi-essen.de/fileadmin/user_upload/RWI/Publikationen/Ruhr_Economic_Papers/REP_23_1004.pdf (Abruf am 18.8.2023).

Noy, Shakked / Zhang, Whitney (2023): Experimental Evidence on the Productivity Effects of Generative Artificial Intelligence. https://economics.mit.edu/sites/default/files/inline-files/Noy_Zhang_1.pdf (Abruf am 18.8.2023).

Nyandwi, Jean (2023): The Transformer Blueprint. A Holistic Guide to the Transformer Neural Network Architecture. In: Deep Learning Revision, 29.7.2023. https://deeprevision.github.io/posts/001-transformer/ (Abruf am 18.8.2023).

OpenAI (2023a): GPT-4 Technical Report. https://arxiv.org/abs/2303.08774 (Abruf am 18.8.2023)

OpenAI (2023b): Our vision fort the future of AGI. https://openai.com/about (Abruf am 18.8.2023).

Oremus, Will (2023): AI chatbots lose money every time you use them. That is a problem. In: The Washington Post, 5.6.2023. www.washingtonpost.com/technology/2023/06/05/chatgpt-hidden-cost-gpu-compute/ (Abruf am 18.8.2023).

Owen, Malcolm (2023): Google Bard: Adequate, but Microsoft Bing blows it away. In: Apple Insider, 18.4.2023. https://appleinsider.com/articles/23/04/18/google-bard-adequate-but-microsoft-bing-blows-it-away (Abruf am 18.8.2023).

Paolillo, Antonio / Colella, Fabrizio / Nosengo, Nicola / Schiano, Fabrizio / Stewart, William / Zambrano, Davide / Chappuis, Isabelle / Lalive, Rafael / Floreano, Dario (2022): How to compete with robots by assessing job automation risks and resilient alternatives. In: Science robotics 7, H. 65. DOI: 10.1126/scirobotics.abg5561 (Abruf am 18.8.2023).

Patel, Dylan / Ahmad, Afzal (2023): Google “We Have No Moat, And Neither Does OpenAI”. In: SemiAnalysis, 4.5.2023. www.semianalysis.com/p/google-we-have-no-moat-and-neither (Abruf am 18.8.2023).

Perrigo, Billy (2023a): Exclusive: OpenAI Used Kenyan Workers on Less Than $2 Per Hour to Make ChatGPT Less Toxic. In: Time Magazine, 18.1.2023. https://time.com/6247678/openai-chatgpt-kenya-workers/ (Abruf am 18.8.2023).

Perrigo, Billy (2023b): 150 African Workers for ChatGPT, TikTok and Facebook Vote to Unionize at Landmark Nairobi Meeting. In: Time Magazine, 1.5.2023. https://time.com/6275995/chatgpt-facebook-african-workers-union/ (Abruf am 18.8.2023).

Pierce, David (2023): Google's Bard chatbot doesn't love me – but it's still pretty weird. In: The Verge, 21.3.2023. www.theverge.com/2023/3/21/23650472/google-bard-ai-chatbot-hands-on-test (Abruf am 18.8.2023).

Pirate Wires (2023): The Boy who cried extinction. https://twitter.com/PirateWires/status/1664296288404242434 (Abruf am 18.8.2023).

Polanyi, Michael (1966): The Tacit Dimension, Chicago: University of Chicago Press.

Reijers, Wessel / Maschewski, Felix / Nosthoff, Anna-Verena (2023): Technosophistische Schattenspiele. In: Philosophie Magazin. www.philomag.de/artikel/technosophistische-schattenspiele (Abruf am 18.8.2023).

Rifkin, Jeremy (1995): The End of Work. The Decline of the Global Labor Force and the Dawn of the Post-Market Era, New York: Putnam.

Rinta-Kahila, Tapani / Penttinen, Esko / Salovaara, Antti / Soliman, Wael (2018): Consequences of Discontinuing Knowledge Work Automation – Surfacing of Deskilling Effects and Methods of Recovery. In: Bui, Tung (Hrsg.): Proceedings of the 51 st Hawaii International Conference on System Sciences. DOI: 10.24251/HICSS. 2018.654 (Abruf am 18.8.2023).

Roose, Kevin (2023a): Inside the White-Hot Center of A. I. Doomerism. In: The New York Times, 11.7.2023. www.nytimes.com/2023/07/11/technology/anthropic-ai-claude-chatbot.html (Abruf am 18.8.2023).

Roose, Kevin (2023b): Aided by A. I. Language Models, Google's Robots Are Getting Smart. In: The New York Times, 28.7.2023. www.nytimes.com/2023/07/28/technology/google-robots-ai.html (Abruf am 5.9.2023).

Ropek, Lucas (2023): After the Death of BuzzFeed News, Journalists Should Treat AI as an Existential Threat. In: Gizmodo, 27.4.2023. https://gizmodo.com/chatgpt-ai-buzzfeed-news-journalism-existential-threat-1849869364 (Abruf am 18.8.2023).

Roser, Max (2022): The brief history of artificial intelligence: The world has changed fast – what might be next? In: Our World in Data, 6.12.2022. https://ourworldindata.org/brief-history-of-ai (Abruf am 18.8.2023).

Saeed, Waddah / Omlin, Christian (2023): Explainable AI (XAI): A systematic meta-survey of current challenges and future opportunities. In: Knowledge-Based Systems 263. DOI: 10.1016/j.knosys.2023.110273. (Abruf am 18.8.2023).

Said, Chris (2023): Double descent in human learning. https://chris-said.io/2023/04/21/double-descent-in-human-learning/ (Abruf am 18.8.2023).

Saquib (2023): Claude AI Chatbot By Anthropic: What Is It? In: New Vision Blog, 5.4.2023. www.newvisiontheatres.com/google-backs-claude-ai-chatbot-by-anthropic (Abruf am 18.8.2023).

Schaeffer, Rylan / Miranda, Brando / Koyejo, Sanmi (2023): Are Emergent Abilities of Large Language Models a Mirage? https://arxiv.org/abs/2304.15004 (Abruf am 18.8.2023).

Schaeffer, Rylan / Khona, Mikail / Robertson, Zachary / Boopathy, Akhilan / Pistunova, Kateryna / Rocks, Jason W. / Fiete, Ila Rani / Koyejo, Oluwasanmi (2023): Double Descent Demystified: Identifying, Interpreting & Ablating the Sources of a Deep Learning Puzzle. https://arxiv.org/pdf/2303.14151 (Abruf am 18.8.2023).

Schaul, Kevin / Chen, Szu Yu / Tiku, Nitasha (2023): Inside the secret list of websites that make AI like ChatGPT sound smart. In: The Washington Post, 19.4.2023. www.washingtonpost.com/technology/interactive/2023/ai-chatbot-learning/ (Abruf am 18.8.2023).

Schloss, David Frederick (1891): Why Working-Men Dislike Piece-Work. In: The Economic Review 1, H. 3, S. 312–326.

Searle, John (1980): Minds, Brains and Programs. In: Behavioral and Brain Sciences 3, H. 3, S. 417–457. https://web-archive.southampton.ac.uk/cogprints.org/7150/ (Abruf am 18.8.2023).

Seemann, Michael (2014): Das neue Spiel. Strategien für die Welt nach dem digitalen Kontrollverlust, Freiburg im Breisgau: Orange-Press.

Seemann, Michael (2023): Danke für den Fisch! In: ctrl-verlust. 14.4.2023. www.ctrl-verlust.net/danke-fuer-den-fisch/ (Abruf am 18.8.2023).

Semuels, Alana (2020): Millions of Americans Have Lost Jobs in the Pandemic – And Robots and AI Are Replacing Them Faster Than Ever. In: Time Magazine, 6.8.2020. https://time.com/5876604/machines-jobs-coronavirus/ (Abruf am 18.8.2023).

Serrano, Luis (2023): What Are Transformer Models and How Do They Work? In: Context by Cohere, 12.4.2023. https://txt.cohere.com/what-are-transformer-models/ (Abruf am 18.8.2023).

Sha, Arjun (2023): Google PaLM 2 AI Model: Everything You Need to Know. In: Beebom, 15.5.2023. https://beebom.com/google-palm-2-ai-model/ (Abruf am 18.8.2023)

Shanahan, Murray (2022): Talking About Large Language Models. https://arxiv.org/abs/2212.03551 (Abruf am 18.8.2023).

Sharro, Karl (2023): “Humans doing the hard jobs on minimum wage while the robots write poetry and paint is not the future I wanted”. Twitter, 15.5.2023. https://twitter.com/KarlreMarks/status/1658028017921261569?s=20 (Abruf am 18.8.2023).

Shumailov, Ilia / Shumaylov, Zakhar / Zhao, Yiren / Gal, Yarin / Papernot, Nicolas / Anderson, Ross (2023): The Curse of Recursion: Training on Generated Data Makes Models Forget. https://arxiv.org/abs/2305.17493 (Abruf am 18.8.2023).

Singhal, Karan / Tu, Tao / Gottweis, Juraj / Sayres, Rory / Wulczyn, Ellery / Le Hou/Clark, Kevin / Pfohl, Stephen / Cole-Lewis, Heather / Neal, Darlene / Schaekermann, Mike / Wang, Amy / Amin, Mohamed / Lachgar, Sami / Mansfield, Philip / Prakash, Sushant / Green, Bradley / Dominowska, Ewa / Arcas, Blaise Aguera y / Tomasev, Nenad / Liu, Yun / Wong, Renee / Semturs, Christopher / Mahdavi, S. Sara / Barral, Joelle / Webster, Dale / Corrado, Greg S. / Matias, Yossi / Azizi, Shekoofeh / Karthikesalingam, Alan / Natarajan, Vivek (2023): Towards Expert-Level Medical Question Answering with Large Language Models. https://arxiv.org/abs/2305.09617 (Abruf am 18.8.2023).

Sorge, Nils-Viktor (2016): „In zehn Jahren werden keine Lkw-Fahrer mehr benötigt“. In: Der Spiegel, 19.9.2016. www.spiegel.de/wirtschaft/unternehmen/autonome-lkw-in-zehn-jahren-werden-keine-fahrer-mehr-benoetigt-a-1112566.html (Abruf am 18.8.2023).

Srnicek, Nick (2022): Data, Compute, Labor. In: Graham, Mark / Ferrari, Fabian (Hrsg.): Digital Work in the Planetary Market. The MIT Press, S. 241–262. DOI: 10.7551/mitpress/13835.003.0019 (Abruf am 18.8.2023).

Stanford AI Index Report (2023): Measuring trends in Artificial Intelligence. https://aiindex.stanford.edu/report/ (Abruf am 18.8.2023).

Steinhardt, Jacob (2023): Emergent Deception and Emergent Optimization. In: Bounded Regret, 19.2.2023. https://bounded-regret.ghost.io/emergent-deception-optimization/ (Abruf 6.8.2023).

Strassberg, Daniel (2023): Wie wir zu Maschinen werden. In: Republik, 20.6.2023. www.republik.ch/2023/06/20/strassberg-wie-wir-zu-maschinen-werden (Abruf am 18.8.2023).

Strickland, Eliza / Zorpette, Glenn (2023): The AI Apocalypse: A Scorecard. How worried are top AI experts about the threat posed by large language models like GPT-4? In: IEEE Spectrum, 21.6.2023. https://spectrum.ieee.org/artificial-general-intelligence (Abruf am 18.8.2023).

Strubell, Emma / Ganesh, Ananya / McCallum, Andrew (2019): Energy and Policy Considerations for Deep Learning in NLP. https://arxiv.org/abs/1906.02243 (Abruf am 11.8.2023).

Sutton, Richard (2019): The Bitter Lesson. www.incompleteideas.net/IncIdeas/BitterLesson.html (Abruf am 18.8.2023).

Touvron, Hugo / Lavril, Thibaut / Izacard, Gautier / Martinet, Xavier / Lachaux, Marie-Anne / Lacroix, Timothée / Rozière, Baptiste / Goyal, Naman / Hambro, Eric / Azhar, Faisal / Rodriguez, Aurelien / Joulin, Armand / Grave, Edouard / Lample, Guillaume (2023): LLaMA: Open and Efficient Foundation Language Models. https://arxiv.org/abs/2302.13971 (Abruf am 18.8.2023).

Turing, Alan M. (1950): Computing Machinery and Intelligence. In: Mind 59, S. 433–460. https://archive.org/details/MIND--COMPUTING-MACHINERY-AND-INTELLIGENCE (Abruf am 18.8.2023).

Underwood, Ted (2023): The Empirical Triumph of Theory. In: In the Moment, 29.6.2023. https://criting.wordpress.com/2023/06/29/the-empirical-triumph-of-theory/ (Abruf am 18.8.2023).

Urban, Elisabeth (2023): ChatGPT ersetzt erste Jobs: Diese Bereiche sind besonders betroffen. In: t3n Magazin, 5.3.2023. https://t3n.de/news/erste-jobs-ersetzt-chatgpt-arbeitsmarkt-betroffene-bereiche-1537447/ (Abruf am 18.8.2023).

Vaswani, Ashish / Shazeer, Noam / Parmar, Niki / Uszkoreit, Jakob / Jones, Llion / Gomez, Aidan N. / Kaiser, Lukasz / Polosukhin, Illia (2017): Attention Is All You Need. https://arxiv.org/abs/1706.03762 (Abruf am 18.8.2023).

Verma, Pranshu / De Vynck, Gerrit (2023): ChatGPT took their jobs. Now they walk dogs and fix air conditioners. In: The Washington Post, 2.6.2023. www.washingtonpost.com/technology/2023/06/02/ai-taking-jobs/ (Abruf am 18.8.2023).

Veselovsky, Veniamin / Ribeiro, Manoel Horta / West, Robert (2023): Artificial Artificial Artificial Intelligence: Crowd Workers Widely Use Large Language Models for Text Production Tasks. https://arxiv.org/abs/2306.07899 (Abruf am 18.8.2023).

Villalobos, Pablo / Sevilla, Jaime / Heim, Lennart / Besiroglu, Tamay / Hobbhahn, Marius / Ho, Anson (2022): Will we run out of data? An analysis of the limits of scaling datasets in Machine Learning. https://arxiv.org/abs/2211.04325 (Abruf am 18.8.2023).

Vincent, James (2016): Twitter taught Microsoft's AI chatbot to be a racist asshole in less than a day. In: The Verge, 24.3.2016. www.theverge.com/2016/3/24/11297050/tay-microsoft-chatbot-racist (Abruf am 18.8.2023).

Vincent, James (2023): Meta's powerful AI language model has leaked online – what happens now? In: The Verge, 8.3.2023. www.theverge.com/2023/3/8/23629362/meta-ai-language-model-llama-leak-online-misuse (Abruf am 18.8.2023).

Virilio, Paul / Lotringer, Sylvère (1984): Der reine Krieg, Berlin: Merve.

Wei, Jason / Tay, Yi / Bommasani, Rishi / Raffel, Colin / Zoph, Barret / Borgeaud, Sebastian / Yogatama, Dani / Bosma, Maarten / Zhou, Denny / Metzler, Donald / Chi, Ed H. / Hashimoto, Tatsunori / Vinyals, Oriol / Liang, Percy / Dean, Jeff / Fedus, William (2022): Emergent Abilities of Large Language Models. https://arxiv.org/abs/2206.07682 (Abruf am 18.8.2023).

Weiser, Benjamin / Schweber, Nate (2023): The ChatGPT Lawyer Explains Himself. In: The New York Times, 8.6.2023. www.nytimes.com/2023/06/08/nyregion/lawyer-chatgpt-sanctions.html (Abruf am 18.8.2023).

Weizenbaum, Joseph (1966): ELIZA – A Computer Program For the Study of Natural Language Communication Between Man and Machine. In: Communications of the ACM 9, H. 1, S. 36–45. https://redirect.cs.umbc.edu/courses/331/papers/eliza.html (Abruf am 18.8.2023).

Wolfram, Stephen (2023): What Is ChatGPT Doing … and Why Does It Work? https://writings.stephenwolfram.com/2023/02/what-is-chatgpt-doing-and-why-does-it-work/ (Abruf am 18.8.2023).

Wolter, Andreas (2023): Deutsche Speditionen ausgebremst. In: Tagesschau, 30.5.2023. www.tagesschau.de/wirtschaft/unternehmen/speditionen-lkw-fachkraeftemangel-100.html (Abruf am 18.8.2023).

Yao, Shunyu / Yu, Dian / Zhao, Jeffrey / Shafran, Izhak / Griffiths, Thomas L. / Cao, Yuan / Narasimhan, Karthik (2023): Tree of Thoughts: Deliberate Problem Solving with Large Language Models. https://arxiv.org/abs/2305.10601 (Abruf am 18.8.2023).

Yeadon, Will / Halliday, Douglas P. (2023): Exploring Durham University Physics exams with Large Language Models. https://arxiv.org/abs/2306.15609 (Abruf am 18.8.2023).

Yudkowsky, Eliezer (2023): Pausing AI Developments Isn't Enough. We Need to Shut it All Down. In: Time, 30.3.2023. https://time.com/6266923/ai-eliezer-yudkowsky-open-letter-not-enough/ (Abruf am 18.8.2023).

Zhou, Viola (2023): AI is already taking video game illustrators' jobs in China. In: Rest of World, 11.4.2023. https://restofworld.org/2023/ai-image-china-video-game-layoffs/ (Abruf am 18.8.2023).

Ziegener, Daniel (2023): Entlassungen bei Axel Springer: Bild-Mitarbeiter sollen „durch KI ersetzt werden". In: Golem.de, 19.6.2023. www.golem.de/news/entlassungen-bei-axel-springer-bild-mitarbeiter-sollen-durch-ki-ersetzt-werden-2306-175083.html (Abruf am 18.8.2023).

Zvi (2023): Microsoft Research Paper Claims Sparks of Artificial Intelligence in GPT-4. In: Lesswrong, 24.3.2023. www.lesswrong.com/posts/FinfRNLMfbq5ESxB9/microsoft-research-paper-claims-sparks-of-artificial (Abruf am 18.8.2023).

Autor

Dr. Michael Seemann studierte Angewandte Kulturwissenschaft in Lüneburg und promovierte 2021 in den Medienwissenschaften an der Universität Tübingen. Er forscht seit 2010 zu Internet und Gesellschaft in verschiedenen Kontexten. Von ihm sind erschienen: „Das Neue Spiel. Strategien für die Welt nach dem digitalen Kontrollverlust“ (2014) und „Die Macht der Plattformen. Politik in Zeiten der Internetgiganten“ (2021).
Er unterrichtet verschiedene Seminare an der Universität zu Köln, der Universität der Künste in Berlin und der Leuphana Universität Lüneburg. 2016 war er als Sachverständiger zum Thema Plattformregulierung im Bundestag. Er hält Vorträge zu den Themen Internetkultur, Plattformen, Künstliche Intelligenz und die Krise der Institutionen in Zeiten des digitalen Kontrollverlusts. 2018 gründete er mit Gleichgesinnten zusammen das Otherwise Network, in dessen Vorstand er seitdem tätig ist.

ISSN 2509-2359