Über die Zukunft einer durch KI geprägten Welt - aus algorithmischer Sicht

Mit Anmerkungen

Von Ulrich Trottenberg (schwarz), Bernhard Thomas (blau)

01-12-2025

Vor rund 90 Jahren hat Bertrand Russell, eins der überragenden mathematischen, literarischen und politischen Genies des letzten Jahrhunderts, sein Buch „Lob des Müßiggangs“ publiziert. Er beschreibt darin eine (fiktive) Welt, in der die Menschen bei weitem nicht mehr so viel wie damals üblich arbeiten müssen. Die Menschen sind in dieser neuen Welt allerdings nicht arbeits-los, sondern (teilweise) arbeits-frei. Was ihre Arbeitsbelastung angeht, können sich die meisten Menschen der Russell-Welt der Schönheit und der kulturellen Vielfalt dieser Welt widmen.

Das ist natürlich eine Utopie, gegen die sich Vielerlei einwenden lässt. Neue Technologien werden stets für mehr Effizienz, Leistungsfähigkeit und Profit eingesetzt, bei gleicher oder gesteigerter Arbeitslast. Und sich „arbeits-frei“ dem Schönen und Kulturellen in dieser Welt zu widmen, basiert wohl auf einem idealisierten Bild vom „Menschen“. Arbeit passt sich immer auch an die Technologien an; man kann sagen, sie wird technologiespezifisch algorithmisiert. Man denke an die „EDV-gestützten“ Prozesse, die seit den Anfängen der Informationstechnik die ständige „Transformation der Unternehmen“ bestimmen, und eigentlich – mit Blick auf Russell - „EDV-*getriebene* Transformation der Arbeit“ bedeutet. Angefangen von frugaler Business-Software bis zum heutigen Stand der großen generativen KI-Anwendungen.

Stehen wir mit der „KI-Revolution“ am Anfang einer solchen neuen Zeit? In der es den Menschen dank KI, dank Robotik und anderer KI-basierter oder KI-naher Umwälzungen, die wir heute noch nicht kennen, viel besser und immer besser geht?

Mit mehr Zeit für „Müßiggang“, weil KI ein Gutteil unserer Arbeit übernimmt – und, zugegeben, auch ein Gutteil des „Müßiggangs“!

Oder stürzt uns die KI-Zukunft in ein nicht mehr beherrschbares Chaos, in dem KI die Regie übernimmt? Wie sieht die von KI geprägte Welt in 10, in 20 Jahren, und wie in der längerfristigen Zukunft aus?

Solcherlei Fragen treiben uns um, Antworten sind schwierig, wenn sie nicht „aus der Glaskugel“ kommen. Ein Grund dafür ist die Komplexität des algorithmischen Geschehens in KI-Systemen im Zusammenspiel mit den menschlichen „Gestaltern“, die bestimmen, welche Daten wie in die Systeme einfließen und welche Ergebnisse als gut, richtig oder zulässig gelten sollen. Das „Chaos“ ist, wenn nicht „beherrschbar“ zum Teil immerhin (statistisch) berechenbar, beherrschbar im Sinne von steuerbar, grundsätzlich. Nicht beherrschbar, im Sinne von nicht vorhersehbar und beeinflussbar,

wird das Ganze durch singuläre Eingriffe von Menschen, sei es aufgrund von Macht, finanziellen Mitteln oder Gier, vor dem Hintergrund einer generellen „Verdummung durch Müßiggang“ (Deskilling). Chaos kann uns auch drohen als Effekt positiver Rückkopplung, d.h. indem KI-Systeme, speziell generative, ihre eigenen Ergebnisse zurückfüttern als Trainingsobjekte (Daten) und Steuerungen (Prompts, Decisions, Agents). Und, dass Menschen zunehmend ihrem „digitalen KI-Zwilling“ vertrauen, anstatt die eigene Urteilsfähigkeit zu bemühen. Aber davon mehr weiter unten.

Sind 10, 20 Jahre oder mehr überhaupt überschaubar? Weitreichende technologische Entwicklungen – zumindest seit dem Internetzeitalter – scheinen nicht mehr linear mit der Zeit zu wachsen, sondern folgen in immer kürzeren Abständen auf der Zeitachse. Lassen wir nur mal die Abfolge von globalen Technologiesprüngen wie World Wide Web, Suchmaschinen, Internet-Business, Social Media, Mobile Networks, Smartphones und mobile Apps Revue passieren und schauen darauf, was sich mit dem Outing von GPT-Modellen, seit Anfang 2023 daraus entwickelt hat. Nun gut, das ist ja gerade unser Thema hier.

Dass KI die Arbeitswelt fundamental verändern wird, das ist absehbar. Viele Arbeitsplätze werden durch „die KI“ ersetzt werden, essentiell beeinflusst werden fast alle Arbeitsplätze.

Und wir reden hier nicht über schön formulierte E-mails oder Geschäftsbriefe, über perfekte Bewerbungen – die ebenfalls wieder von KI-Systemen „gelesen“, verglichen, bewertet und beantwortet werden. Sondern über alle Prozesse in Wirtschaft, Verwaltung, Kultur, Wissenschaft und Politik, die essentiell auf multimodalen Informationsfluss und Kommunikation angewiesen sind. Arbeit - und damit die Arbeitsplätze - wird nicht verschwinden aber produktiver, effizienter werden müssen, gestützt oder getrieben von KI-Prozessen („Agentic AI“). Für SW-Entwickler gibt es beispielsweise zurzeit schon Qualifikationsprozesse für eine KI-basierte SW-Entwicklung mit Einhaltung aller geforderten SW-Standards. Es gibt kaum einen Bereich, der sich nicht schon mit den Möglichkeiten bzw. Notwendigkeiten auseinandersetzt. Physische Abläufe (wie in Handwerk, Produktion, Handel) werden noch auf die entsprechende Robotik und embedded AI warten können („der künstliche Dachdecker“). Ihre Optimierung und Steuerung, seit langem schon algorithmisiert und IT-gestützt, ist aber ein fruchtbares Feld für die moderne KI. (NB: Nicht, dass wir nicht schon lange vor der öffentlichen Wahrnehmung von ChatGPT und Co Unternehmen und Bildungseinrichtungen auf die Notwendigkeit hingewiesen hätten, sich mit KI zu beschäftigen.)

Viele junge Menschen stellen sich dementsprechend schon jetzt auf die erwartbare KI-Zukunft ein und bemühen sich um ein Verständnis der KI. Viele Schulen stehen der KI-Zukunft der Menschheit heute noch ziemlich ratlos gegenüber.

Ratlos stehen Schulen, und auch Hochschulen, der KI seit Anfang 2023 primär in Bezug auf die Möglichkeiten von LLMs gegenüber. Ihr Problem ist primär, dass Schüler und Studierende mittels LLMs Aufgaben aus allen Disziplinen einfach bearbeiten können, indem sie Ergebnisse per LLM (und geeignetem Prompt) produzieren können und damit u.a. das Problem einer gerechten Beurteilung bzw. Benotung entsteht. Eine durchaus gute, 60-seitige Bachelor-Arbeit zu einem hochspezialisierten Thema ist selbst im naturwissenschaftlichen Umfeld mit geeigneter „Programmierung" (durch Prompts) der LLMs zu generieren.

Dabei wird ein zweiter wichtiger Aspekt nach wie vor vernachlässigt: Das Verständnis dafür, was „KI" ist, wie es zu der „Intelligenz" kommt, welche (algorithmischen) Verfahren dem „Machine Learning" zugrunde liegen und wie man sie verstehen kann. (Dass das durchaus schon auf Unterstufen-Niveau möglich ist, haben wir bereits Anfang der 20er Jahre mit unseren „KI-Stories" und im Schulexperiment gezeigt.)

Und wie es mit der KI weitergeht, ist ja wirklich eine der ganz großen Fragen.

Müssen wir regelmäßig (oder immer mal wieder?) im KI-Bereich mit disruptiven Entwicklungssprüngen rechnen, wie wir sie Ende 2022 erlebt haben, als ChatGPT plötzlich allgemein verfügbar wurde. Die großen Sprachmodelle (LLMs) haben seitdem die Welt „kommunikativ" verändert und bieten mit geeigneten Prompts ungeahnte sprachliche Möglichkeiten. ChatGPT war also Ende 2022 in der Anwendung und Nutzung ein ungeheurer Sprung, nicht aber in der Methodik. Denn die GPT-Methoden (generative pretrained transformer) waren seit Jahren bekannt, waren veröffentlicht und wurden von Experten intensiv benutzt.

Um diese Frage anzugehen, muss man sich fragen, was diesen disruptiven Sprung ausgelöst, ermöglicht hat. Wenn, wie so oft zuvor, die Idee, die Methode, das Potenzial dazu bereits latent vorhanden war. In diesem Fall kamen die massive verfügbare Rechenpower (GPU-Farmen) und der Zugang zu den immensen Mengen an Text- und Bilddaten im Internet zusammen. Und natürlich auch die „Skalierung" der grundlegenden Algorithmen, die durch die Rechnerressourcen möglich wurde. Auf dieser Basis und der Bereitschaft eines Unternehmens den immensen Aufwand des Trainings und Re-Trainings mit menschlichem Feedback zu treiben, wurde das Ganze „sprungbereit". Der Sprung selbst vollzog sich autokatalytisch aufgrund der allgemeinen Öffnung und der natürlich-sprachlichen Interaktion mit dem System. Die Frage für die Zukunft ist also: sehen wir Neues mit ähnlichen Charakteristika aus der KI auf uns zukommen? Mehr dazu am Ende.

Sehr viele Menschen haben, gerade dank der großen Sprachmodelle wie ChatGPT, mittlerweile eine Ahnung, was KI bedeuten könnte. Viele sind fasziniert, viele sind aber auch verunsichert. Viel mehr wissen die meisten Menschen nicht. Sie kennen das Wort „Algorithmus", weil es in den Medien täglich benutzt wird. Aber was das ist, ein

„Algorithmus“, und welche Bedeutung Algorithmen für die KI haben, das weiß allenfalls eine Minderheit der Menschen.

Die Menschen bzw. Nutzer sind fasziniert von KI in Form von den „generativen“ Fähigkeiten von LLMs – nicht nur vom Inhalt, sondern vor allem von der sprachlichen und medialen Qualität des Outputs, die unwillkürlich als menschen-ähnlich wahrgenommen wird. Und man versteht nicht, wie das dazu kommt – solange man nicht eine Vorstellung von den zugrunde liegenden Abläufen (Algorithmen) entwickelt hat.

Also: Was ist diese KI überhaupt? Und was ist ein Algorithmus für ein Gebilde?

Als Algorithmiker sollte ich das klar beantworten können, sollte man meinen. Aber ein Algorithmus ist heute nicht genau dasselbe wie vor 10 Jahren. Vor 10 Jahren hätte ich zum Beispiel „definiert“: „Ein Algorithmus ist eine eindeutige Handlungsanweisung, mit der eine wohldefinierte Aufgabe in endlich vielen Schritten gelöst wird.“

In dieser Hinsicht, sind die meisten heutigen Arbeitsumgebungen seit langem bereits „algorithmisiert“, die Mitarbeitenden in klare, auf Informationsflüsse basierende Prozesse eingebunden. Business Process Engineering und IT-Systeme – heute auch als „Digitalisierung“ verstanden – haben die Arbeitswelt von jeher verändert, mit dem Ziel Produktivitäts- und Effizienz-Verbesserung durch „eindeutige Handlungsanweisung, mit der eine wohldefinierte Aufgabe (IT-geführt) in endlich vielen Schritten gelöst wird“. Was kommt also Neues hinzu mit KI, und mit generativer KI insbesondere?

Heute gibt es Algorithmen, von denen man zumindest in Frage stellen kann, ob diese Definition noch exakt zutrifft: Diese neuen „lernenden“ Algorithmen verändern sich selbst, passen sich an neue Situationen an.

Auf solchen Algorithmen basiert KI, gleich welcher Art. Ob Erkennen und Klassifizieren von Situationen, Erstellen von Vorhersagen oder eben die Generierung von multi-medialen Inhalten. Und „die KI“ – die Algorithmen und Anwendungen - entwickelt sich rasant weiter. Dabei muss man sich klarmachen, wer was weiterentwickelt. Sind es die SW-Schmieden, die Tech-Konzerne, also menschliche Entwickler? Sind es - indirekt - die Massen an Nutzern der LLMs? Die KI kann sich – in Ansätzen schon heute – durchaus selbst „weiterentwickeln“. Wohin weiterentwickeln, ist eine Frage der Zielsetzung und damit, wer oder was Zielgeber ist. Wir müssen unterscheiden und entscheiden, ob „die KI“ Objekt oder Subjekt der Entwicklung sein soll.

Was also wissen KI-Algorithmiker mehr über die Zukunft der KI-geprägten Welt als Publizisten, Politiker, interessierte Laien?

Man kann erwarten, dass Fachleute, die sich mit Algorithmen auskennen, die Algorithmik von KI-Systemen zumindest prinzipiell verstehen, sie durchdrungen haben, und sei es, weil sie sie selbst konstruiert haben. Da haben sie sicher einen Vorteil. Daraus erwächst aber auch die Verpflichtung, der Auftrag, Nicht-Fachleuten

verstehbare und sachlich korrekte Erklärungen zu geben, auf deren Basis z.B. Verantwortliche in Politik, Bildung und Medien zuverlässig gesellschaftliche Verantwortung für KI übernehmen können. Interessanterweise kann man solche Erklärungen inzwischen ohne weiteres von LLMs erfragen.

Die KI, wie wir sie heute und auf absehbare Zeit kennen, ist ein algorithmisches, also mathematisches Konstrukt. Sie ist absolut nichts Geheimnisvolles, nichts, was selbständig denkt, nichts, was einen Willen hätte, sie ist Mathematik, und - für viele Menschen vielleicht überraschend - vergleichsweise simple Mathematik: ein bisschen Statistik, ein bisschen Numerik und Optimierung, ein bisschen Approximation. Wenig Theorie, dafür umso mehr experimentelle Mathematik, sehr viel algorithmisches Ausprobieren.

„Die KI" wird umgangssprachlich mit Begriffen beschrieben, die als Analogien brauchbar scheinen aber algorithmisch keinen Sinn machen. Oft auch in den Medien mit Aufmerksamkeitsbedarf. Selbst Fachleute der KI-Szene prägen ständig neue Begriffe, die eine nicht-technische, d.h. umgangssprachliche Bedeutung provozieren. Wenn man den algorithmischen Boden verlässt, gerät man schnell in die Denk- und Gefühlskategorien, die eher das Feld der Kognitionswissenschaften, der Philosophie, Psychologie etc. sind.

Wenn die KI nichts anderes als „simple Mathematik" ist, wodurch erreicht sie dann die ungeheure Wucht, mit der sie uns so beeindruckt und immer wieder überrascht? Es sind die Billionen von Daten und die Milliarden von Parametern, mit denen die mathematisch simplen KI-Algorithmen „souverän" umgehen.

Technisch ist der „Lernprozess", den diese mathematischen Algorithmen durchlaufen – und nur deshalb bezeichnet man sie als „intelligent" - die Anpassung dieser vielen Parameter durch eine mathematische Form von Trial-and-Error (und Korrekturen), so dass dadurch die Vorgabedaten gut repräsentiert werden. Da es sich hierbei aber um immense Größenordnungen handelt, ist es aussichtslos verstehen zu wollen, welche Auswirkungen einzelne Parameterwerte oder Input-Daten im Algorithmus haben. Erst recht, wenn das „Gelernte" für neue Anfragen oder Aufgaben angewendet werden soll. (Mehr dazu unten unter „Explainable AI").

Ist die Tatsache, dass wir es bei der KI letztlich „nur" mit (simpler) Mathematik zu tun haben, beruhigend im Sinne von „was kann simple Mathematik schon anrichten?" Nicht unbedingt.

Denn Mathematik ist völlig wertneutral, kann für alles eingesetzt werden, für die größten Errungenschaften und für die größten Katastrophen, für die Optimierung unserer Lebensbedingungen und die Optimierung von Vernichtungskriegen.

Was kann die Algorithmik zu den KI-Perspektiven sagen, hier und jetzt und in der Zukunft weltweit? Kann sie überhaupt etwas zu den Perspektiven sagen? Müssen wir nicht damit

rechnen, dass die lernenden Algorithmen ohnehin alles übernehmen und sich selbst unkontrolliert und unkontrollierbar weiterentwickeln? Und die Menschheit algorithmisch auf der Strecke bleibt?

Gehen wir kurz auf diese Fragen ein. Wie schon zuvor erwähnt, ist dies ein Frage der Zielsetzung und wer die Ziele setzt. Der bisherige (philosophische) Konsens ist, dass KI-Systeme sich keine eigenständigen Ziele, oder Probleme, stellen können. Ihr inhärentes, eingebautes Ziel ist, technisch gesehen, die mathematische Optimierung ihrer Outputs bezüglich *vorgegebener* Ziele. Im Dialog-Kontext gesehen bedeutet das, „die KI" will dem Nutzer „gefallen", etwas Gutes tun, ihm helfen. Über menschliches Feedback, technisch bestärkendes Lernen (mit Menschen-Hilfe), wird die KI dafür „belohnt" (oder bestraft). Der Algorithmiker erkennt dies als menschliche Interpretation der (numerischen) Optimierung. Viele LLMs stellen sich am Ende ihrer Ausführungen zu einem Prompt *scheinbar* weitere, eigene Probleme. Aber auch das sind lediglich Fortsetzungsmöglichkeiten des Chats, in Frageform oder Angebot formuliert, um den Nutzer im Chat zu binden. Das überrascht nicht wirklich. Und oft genug kann das LLM seine Angebote gar nicht umsetzen.

Die menschliche Einflussnahme auf die Zielsetzung ist dagegen erheblich. Insbesondere durch Hintergrundeinstellungen (System-Prompts), wie das Modell auf Anfragen und Anforderungen reagieren soll. Die Möglichkeiten, ethische, moralische, politische oder sonstige gesellschaftlich relevante Vorgaben (Ziele) zu implementieren bzw. nach zu trainieren liegen entweder in den Händen Weniger oder im Feedback der Massen weltweit. Die Frage der *Kontrollierbarkeit* der lernenden Algorithmen stellt sich auf dieser Ebene, nicht auf der algorithmischen.

Damit müssen wir nicht rechnen.

Um abzuschätzen, womit wir rechnen müssen, blicken wir auf die drei fundamentalen Komponenten der KI:

- auf die Algorithmik, also die Mathematik,
- auf die Daten, auf welche die Algorithmen zugreifen,
- auf die Superrechner, die zur Auswertung der KI benutzt werden.

Dazu stellen wir fest:

1. Die Mathematik ist heute und morgen weitgehend dieselbe. Sie bleibt in der überschaubaren Zukunft prinzipiell gleich, sie verändert sich nicht grundlegend. Es kann sein, dass im Detail vieles besser geht und z.B. die Optimierungsalgorithmen deutlich beschleunigt werden. Und was die „Quanten-Algorithmen" für die KI bedeuten, ob sie zu Durchbrüchen, zu exponentiellen Beschleunigungen führen - das können wir heute noch nicht realistisch abschätzen.
   Mit neuen Algorithmen, die auf künftige Quanten-Computer ausgerichtet sind, hat man heute bereits mit Quanten Machine Learning (QML) Methoden die

algorithmischen Grundlagen gelegt für eine „Quanten-KI“ (nicht zu verwechseln mit „Quantization“ von LLMs, mathematische Verfahren, um Größe und Rechenzeit bei der Anwendung zu reduzieren). Für QML erwartet man, dass Teile der KI-Rechen-prozesse potenziell „exponentiell“ beschleunigt werden können. Die Konsequenzen für eine künftige KI des Alltags sind zurzeit noch nicht realistisch abzuschätzen. Aber die Mathematik bleibt auch in der Quantenwelt prinzipiell die gleiche.

2. Was die Menge der verfügbaren Daten angeht: Wenn wir annehmen, dass das Internet repräsentativ ist für die Menge der verfügbaren Daten, dann ist festzustellen: Das Internet ist erst mal da, und es weitet sich immer weiter aus. Aber - abgesehen von den Anfängen des Internets - verdoppelt sich die Menge der Textdaten nicht regelmäßig von einem zum nächsten Jahr. Sie wächst vielleicht jährlich im Mittel um einen kleinen Prozentsatz - auch das ist exponentielles Wachstum. Aber es ist kein unkontrollierbares, auf keine Weise mehr beherrschbares Wachstum. Ähnliches dürfte für Videodaten, Audiodaten und andere Datentypen gelten.
   Die Datenfülle für und mit KI wächst auch dadurch an, dass durch die immense Nutzung milliardenfach KI-Output generiert wird, der selbst wieder als Daten (Input) verfügbar ist von den Systemen genutzt wird – solange es Speicher und Compute-Ressourcen das zulassen. Schon seit Anfang 2025 übersteigt der generierte Text die Menge der originären Daten im Internet; bei Bild- und Audiodaten zeigt sich die gleiche Entwicklung.
   Die Datenmenge ist der eine Wachstumsaspekt. Ein anderer ist das Verhältnis von gesichert korrekten Daten zu von der KI „erfundenen“ Daten, das entsprechend dazu neigt, exponentiell zu schrumpfen. Zudem entsteht ein selbstverstärkender Nebeneffekt dadurch, dass immer mehr Menschen dazu neigen, die KI für sich „sprechen“ zu lassen, als eine Art „KI-Zwilling“; d.h. neue menschen-generierte neue Daten haben schon „die KI“ durchlaufen.

3. Die Supercomputer werden immer leistungsfähiger und schneller, aber auch bei den Rechnern haben wir es nicht mit völlig unkontrollierbaren exponentiellen Leistungssteigerungen zu tun.

Eine vierte, allerdings nicht ursprünglich algorithmisch fundierte Komponente, mit der wir rechnen müssen, ist die zunehmende Beeinflussung bzw. Zielgebung im Hintergrund der Systeme, wie schon zuvor erwähnt.

Aus diesen Feststellungen können wir schließen, dass durch die Konstanz der Mathematik und durch das „langsame“ exponentielle Wachstum der Datenfülle mit unbeherrschbar sprunghaften Entwicklungen der KI nicht zu rechnen ist. Aus algorithmischer Sicht müssen wir mit plötzlichen disruptiven KI-Katastrophen also nicht rechnen.

Also müssen wir uns aus algorithmischer Sicht keine KI-Sorgen machen?

Ganz so einfach und beruhigend ist die Situation leider doch nicht.

Trotz der mathematischen Durchsichtigkeit der KI lassen sich

- Kontrollverlust und
- Unvorhersehbarkeit

grundsätzlich nicht völlig vermeiden.

Man kann das schon an spielerischen Anwendungen der KI erkennen und demonstrieren: Beim Gospiel, dem wohl komplexesten bekannten Brettspiel, ist die KI seit gut 10 Jahren den besten menschlichen Gospielern überlegen ist, und die KI hat Spielvarianten entdeckt, die zuvor noch kein Mensch als Spielmöglichkeiten erwogen hat.

Das war noch vor dem großen Durchbruch von ChatGPT und anderen LLMs und technisch sehr aufwändig. Aber im Prinzip ähnlich verstehbar. Beim Go wie auch bei Texten und Bildern gibt es stets mehrere Möglichkeiten, den momentanen Stand fortzusetzen, als Spielzug, nächste Phrase, nächstes Bild, von denen der Algorithmus je „Toleranz“ eine zufällig auswählt. Spielt das System im Training gegen sich selbst, können so Spiele entwickeln, und damit Zug-Sequenzen, auf die ein menschlicher Spieler nie gekommen wäre. Bei LLMs spricht man da von „Halluzinieren“ - oder auch von „Kreativität“, je nach Verwendungszweck. KI erfindet Neues.

So hat man heute schon vielfach KI-Nutzer, die mit einer LLM („ihrer KI“) per Dialog eine persönliche oder gar intime Beziehung eingehen, teilweise mit individual-psychologischen Problematiken als Folge. Andererseits kann schon ein vollständiger wissenschaftlicher Prozess mitsamt peer-reviewter Veröffentlichung von KI-Systemen generiert oder begleitet werden. Die „KI“ als Forschungspartner. (s. extensity.ai).

Man muss hier auch XAI (explainable AI) ansprechen, den Bereich der Bemühungen, die mit KI erzielten Ergebnisse in voller Allgemeinheit zu verstehen und zu erklären. Auch in diesem Bereich sind viele Fragen trotz intensiver Bemühungen und bemerkenswerter Erfolge noch offen und ungeklärt.

Die KI-Algorithmen und die (mathematischen) Verfahren, die bei Training und Inferenz (i.e. die Ableitung der Ergebnisse aus neuem Input bei trainierten Systemen) sind komplex aber gut verstanden und erklärbar. Für einen mathematische ausgebildeten Algorithmiker. Warum aber bei einem bestimmten Input, einer Anfrage, einem Prompt, ein ganz bestimmtes, sogar unerwartetes Ergebnis herauskommt, können auch diese nicht oder nur schwer erklären. Das hängt mit der Größe (Milliarden von Parametern) und ein Vielfaches mehr an Berechnungen zusammen, die bei jeder Antwort vom Modell durchgeführt werden. Unter anderem. Es ist auch schwer zu verstehen, welchen Einfluss welche Trainingsdaten auf das konkrete Ergebnis haben. Eine vollständige

Erklärung durch eine geschlossene mathematische Theorie scheint nicht erreichbar. Hier bemüht sich die XAI-Forschung (Explainable AI) um tiefere Erkenntnisse. Dabei versucht man durch vielfältige, systematische Experimente das Verhalten der Modelle zu ordnen und ursächlich zu beschreiben. Das KI-Modell in seiner Komplexität als Organismus und Forschungsobjekt. Idealerweise versucht man, das KI-System so zu gestalten, dass es selbst erklärt, wie es zu einer Einschätzung oder Vorhersage kommt.

Bei den generativen LLMs beobachtet man das „Verhalten" der Modelle. Man „experimentiert" über Systemprompts, die eine Art genereller Grundeinstellung des Modells festlegen, z.B. in ethischer oder politischer Hinsicht. Über User Prompts, eingeschleuste spezifische Dokumente (anstelle der amorphen Internetquellen), Feedback-Schleifen, Chat-Historien, Wiederholungen und Vergleiche verschiedener LLMs. Zudem kann inzwischen jeder interessiert Laie lokal auf seinem PC kleinere Varianten der vorhandenen LLMs selber modifizieren. Verschiedene LLMs haben einen „Nachdenk"-Modus, bei dem sich verfolgen lässt, wie und mit welchen Quellen, mit welchen „Überlegungen" es zum abschließenden Ergebnis kommt. Das sieht nach „Erklärung" aus, ist es aber nicht im Sinne von XAI. Es erklärt nicht, wie der zugrunde liegende LLM-Algorithmus mit seinen Milliarden Parametern gerade zu dieser Herleitung (Berechnung) der Ergebnisse kommt.

Zusammengefasst kann ich als Algorithmiker aber sagen:

Die KI ist Mathematik. Wir können den Missbrauch der Mathematik und der KI nicht ein für alle Mal ausschließen. Aber die Mathematik verselbständigt sich nicht. Die Mathematik und die KI sind Werkzeuge. Die können wir nutzen und kontrollieren.

Oder doch nicht? Die Bausteine und Werkzeuge verselbstständigen sich sicherlich nicht. Aber Gebilde aus diesen Bausteinen? Ist KI nur die Summe ihrer (algorithmischen) Teile? Wir haben oben festgestellt, dass dies nur eine von drei, oder vier, fundamentalen Komponenten ist, die die heutige KI ausmachen. Wir sehen, dass die anderen Komponenten, Datenfülle, Rechen- und Speicherkapazität und auch die Einflussnahme kontinuierlich und erheblich wachsen. (Schwarz Digit plant Compute-Fabriken mit 100.000en GPUs, deren Abwärme allein als Fernheizung genutzt werden kann. LLMs werden, testweise, so „vorgepromptet", dass sie kritische User-Chats direkt an die Polizei melden können.) Auch das ist alles beherrschbar. Aus der Systemtheorie wissen wir aber, dass selbst einfache kontinuierliche Prozesse, im Zusammenwirken, emergente Qualitäten entwickeln können und dabei durchaus auf „Kipppunkte" zusteuern können.

Alles Weitere liegt am Menschen – wie immer bisher.

Anmerkung: Die Texte sind an keiner Stelle mit LLM-Hilfe geschrieben. Dennoch, bei genauer Betrachtung, sind sie aber auf gleichem Wege entstanden: Über lange Zeit

Erlerntes, durch Gespräche und Praxis Erfahrenes, Gelesenes speist die Entscheidung, welcher Satz in welcher Form auf den vorausgehenden folgen soll. Das geht auch im Dialog. Einen Unterschied sehen wir aber doch: Wir wissen um die Bedeutung dessen, was wir schreiben, wir haben verstanden, was wir schreiben – oder zumindest glauben wir das.