Was bedeutet ChatGPT?

Chancen und Herausforderungen für uns Alle

Professor Dr. Alexander Pretschner
Software&Systems Engineering an der TU München
Forschungsinstitut des Freistaats Bayern für software-intensive Systeme
Bayerisches Forschungsinstitut für Digitale Transformation
München, Deutschland

Was bedeutet ChatGPT?
Chancen und Herausforderungen für uns Alle

1. Was kann künstliche Intelligenz wie gut?

ChatGPT[i] und vergleichbare Anwendungen sind Vertreter einer speziellen Form Künstlicher Intelligenz (KI). Die öffentliche Debatte zur KI ist auch jenseits von Anwendungen wir ChatGPT von großer Aufregung geprägt. Auf der einen Seite stehen etwa in Deutschland vielleicht landestypisch die Risiken im Vordergrund, die Sorgen, die Gefahren des Unwägbaren, die German Angst. Andererseits wird jenseits von eher feuilletonistischen Überlegungen dazu, ob Künstliche Intelligenz nun kreativ sei oder nicht, mindestens unterschwellig suggeriert, dass Künstliche Intelligenz eigentlich alle (zumindest technischen) Probleme lösen könne. Besonders auffälligen Niederschlag findet dies in Aussagen etwa von CEOs wie Elon Musk oder Alex Karp,[ii] die KI nicht nur auf dem Schlachtfeld mit der Atombombe vergleichen. Diesem Vergleich wohnt eine bemerkenswerte Diskursverschiebung inne. Anstatt auch angesichts vergangener KI-Winter nüchtern die angemessene Frage zu stellen, was KI in welchem Kontext in welchem Ausmaß zu leisten vermag, ist die unterschwellige Annahme in diesem Framing, dass KI mittlerweile tatsächlich alles kann und nun, als anstehende wichtigere Aufgabe, eingehegt werden muss. Es ist verständlich, dass derartige Dramatisierungen den öffentlichen Diskurs pfeffern und die Beobachtung nicht fernliegend, dass die Urheber solcher Aussagen ihr Geld mit KI verdienen und damit auch Erfolg haben, wie der Kurs des NASDAQ seit Anfang des Jahres zeigt.

Künstliche Intelligenz hat ohne jeden Zweifel viel Potential, und die Fortschritte der letzten Jahre in Theorie, Hardware und Software sind atemberaubend, wie nicht nur der spektakuläre Erfolg von Large Language Models wie (Chat)GPT zeigt. Gleichwohl ist die implizite Annahme, KI «könne alles», selbstverständlich falsch. Dazu ist zunächst zu beobachten, dass die in der Öffentlichkeit präsentierten Erfolge, so spektakulär sie sind, häufig in Anwendungsbereichen liegen, in denen eine objektivierte Messung der Qualität nicht notwendig zu sein scheint: Wie gut sind, um mit einem einfachen Beispiel zu beginnen, die Hintergrundfilter in Videokonferenzsystemen? Wie gut sind Bildsynthesesysteme, die photorealistische Bilder von Astronauten auf Schimmeln erzeugen? Wie gut sind von ChatGPT erzeugte Texte? In diesen Anwendungen kann die Bewertung dessen, was hinreichende Qualität ist, häufig im Nachhinein und eher intuitiv erfolgen: You know it when you see it. Gleichzeitig ist es offensichtlich, dass in medizinischen Anwendungen oder beim Bau autonomer Fahrzeuge vorab Gütekriterien und Schwellwerte festgelegt werden müssen. Und wenn KI in Unternehmen angewendet wird, ohne zu Beginn über objektivierbare Gütekriterien und Schwellwerte nachzudenken, gleichzeitig aber große Erwartungen zu kultivieren, stellt sich nicht selten Ernüchterung ein. Das ist der Fall, wenn im Nachhinein festgestellt wird, dass die Erwartungen nicht erfüllt werden können, weil etwa Daten fehlen, unvollständig oder falsch sind oder das zu lösende Problem sich schlicht als zu schwierig herausstellt.

Ein eindrückliches Beispiel für die Fähigkeiten und Defizite von KI sind Deep Fakes, als die man künstlich erstellte und täuschend echte Videos von Personen bezeichnet. Deep Fakes als solche zu erkennen, sehen viele nicht nur als technische, sondern vor allem auch als große gesellschaftliche Herausforderung. 2020 gab es einen vom damaligen Facebook-Konzern ausgelobten internationalen Wettbewerb, die Deep Fake Detection Challenge.[iii] Mit einem Preisgeld von $1.000.000 wurde die beste Software zur Erkennung von Deep Fakes ausgezeichnet, die in 2/3 der Fälle korrekt erkennen konnte (Accuracy), ob ein Video ein Deep Fakes ist oder nicht. Für einen praktischen Einsatz ist eine solche Genauigkeit vermutlich zu gering. Das Beispiel zeigt indes gleichzeitig eindrücklich, wie leistungsfähig KI ist, nämlich im *Erstellen* von Deep Fakes, und wo gleichzeitig heute noch deutliche Grenzen zu erkennen sind, nämlich im *Erkennen* von Deep Fakes. (Der Leser

und die Leserin frage sich an dieser Stelle, welche Qualität für die Erkennung von Deep Fakes in welchem Kontext denn *hinreichend* gut sei.) Zu erkennen, ob ein Text von ChatGPT oder von einem Menschen erstellt wurde, ist mit überzeugender Qualität heute ebenfalls nicht möglich.

Die zugrundeliegende Schwierigkeit ist konzeptioneller, technischer, pragmatischer und anwendungsbezogener Natur. Zum einen wird Künstliche Intelligenz, vor allem das heute prominente Maschinenlernen, häufig in denjenigen (zahlreichen!) Feldern angewendet, in denen das Problem nicht präzise beschreibbar und deswegen prinzipiell nur schwer regel- oder algorithmenbasiert lösbar ist – wie beschreibt man Regeln zur Erkennung von «Fußgängern», wenn Konzepte wie «Torso» oder «Arm» oder «Regenschirm» ihrerseits erst präzisiert werden müssen? Und wenn das Problem nicht klar fassbar ist, ist es offenbar schwierig, die Güte der Lösung klar zu erfassen. Zweitens sind die Gütekriterien als solche nicht kanonisch und auch nicht eindeutig: Es gibt zahlreiche auf der Konfusionsmatrix basierende Gütemaße, die über gegebene Testdaten empirisch ermittelt werden und insofern ihrerseits von der Qualität und erwünschten oder nicht erwünschen Repräsentativität der Testdaten abhängen. Solche Maße werden heute durch Maße zu Robustheit oder Fairness ergänzt, die intensiv erforscht werden – und für die es jeweils viele unterschiedliche Vorschläge gibt, die etwa im Fall von Fairness nachweisbar nicht gleichzeitig erfüllt werden können. Drittens liegt es in der Natur maschinengelernter Modelle, dass sie aus Beispielen lernen und diese verallgemeinern und insofern Fehler bei der Generalisierung nie vollständig ausgeschlossen werden können. Viertens ist die Verfügbarkeit adäquater, korrekter, vollständiger und markierter («labelled») Daten in hinreichender Menge stark abhängig vom betrachteten Kontext und stellt in der Praxis eine der größten Herausforderungen dar.

Es lässt sich festhalten, dass maschinengelernte Anwendungen in vielen Bereichen tatsächlich spektakuläre Erfolge vorweisen. Solche Anwendungen sind im technischen Sinn selten «gut» oder «schlecht». Manchmal ist eine Objektivierung der Qualität nicht notwendig. Häufig ist diese Frage aber relevant, wenn sie eher auf ein «wie gut ist diese Anwendung?» hinausläuft und, wie weiter unten erläutert, häufig die zusätzliche Frage nach einem «gut genug» zeitigt.

2. Anwendungen von ChatGPT und unsere Erwartungen

ChatGPT[1] und verwandte Produkte oder Technologien, deren Funktionsweise und zugrundeliegende Technik wir in Abschnitt 4 kurz skizzieren werden, sind Systeme, die auf Basis einer Anfrage, dem sogenannten *Prompt*, Texte erzeugen können. Auf die Frage «Was hat Albert Einstein in Wien gemacht?» gibt ChatGPT einen verblüffend gut formulierten Text aus, der in Teilen faktisch korrekt und in Teilen inkorrekt ist. Die Qualität der Antwort hängt stark von der Art ab, wie die Anfrage formuliert ist, was das neue Betätigungsfeld des «Prompt Engineering» aus der Taufe gehoben hat.[iv] Warum unabhängig von der Beschaffenheit des Prompts Korrektheit prinzipiell nicht garantiert werden kann, erläutern wir ebenfalls in Abschnitt 4.

Als System für die Erstellung von Texten hat ChatGPT naturgemäß zahlreiche Anwendungen. Die Idee ist dabei stets, dass aus einem kurzen Prompt ein längerer Text erstellt wird. Dieser Text kann entweder neue «Ideen» enthalten und so der Inspiration dienen; er kann aber auch eine Ausformulierung von Stichpunkten sein oder umgekehrt eine Zusammenfassung eines längeren Textes. ChatGPT kann dementsprechend produktivitätssteigernd wirken, wenn ein Mensch nicht mehr einen kompletten Text formuliert, sondern stattdessen die wesentlichen Inhalte oder die Argumentationslinie vorgibt und dann den erzeugten Text noch redigiert. Beispielhaft seien hier die Erstellung von Emails oder Social Media-Posts genannt; die Erstellung von Textvorlagen in der öffentlichen Verwal-

[1] ChatGPT wird hier stellvertretend für sog. Large Language Models im Allgemeinen und auch spezielle LLMs wie GPT4 oder PaLM oder LLaMa verwendet. Wir erläutern die Konzepte in Abschnitt 4, verzichten zunächst aber auf eine saubere Unterscheidung zwischen GPT und ChatGPT.

tung,[v] Arztbriefen oder Klageschriften; von Reden, Grußworten, Prüfungsfragen oder Gutachten; von Programmcode und Excel-Formeln; von Visualisierungsideen, Bildern und kompletten Videos; die Erstellung oder Optimierung von Drehbüchern; das Straffen von Texten sowie die Erstellung von Zusammenfassungen,[vi] Übersetzungen, Dokumentationen und automatisierten Reaktionen auf Anfragen im Call Center.

Da ChatGPT über die Fähigkeit verfügt, Informationen zum Kontext aus vorangegangenen Interaktionen in neue Interaktionen einfließen zu lassen, eignet es sich auch als Dialogpartner. Der Verfasser dieses Textes verwendet ChatGPT u.a. dazu, eigene neue Ideen im Frage stellen zu lassen. Die Dialogfähigkeit ist ein mächtiges Werkzeug auch dann, wenn ChatGPT zur (manchmal falschen!) Erläuterung von Sachverhalten oder Programmcode verwendet wird, was Potential insbesondere in der Lehre von der Primarstufe über die Universität bis hin zur beruflichen Weiterbildung bietet.

Besonders attraktiv erscheint ChatGPT auch, wenn es als natürlichsprachliche Schnittstelle zu Systemen fungiert, mit denen bisher nicht oder nur eingeschränkt natürlichsprachlich kommuniziert werden konnte. Wenn (technisch subtile) Anfragen an Datenbanken nun synthetisiert werden können oder das gesamte dokumentierte Wissen einer Organisation zum Training ChatGPT-artiger Technologie verwendet wird und dann natürlichsprachliche Anfragen über dieses gesamte, heute oft nicht zugreifbare Wissen, beantwortet werden können, erscheint das Potential wahrhaft gigantisch.

Allen Anwendungen ist gleichwohl gemein, dass die generierten Artefakte vom Menschen überprüft werden müssen, weil die Technik falsche oder irreführende Antworten derzeit nicht ausschließen kann – und falsche oder irreführende Antworten auch nicht eine seltene Ausnahme darstellen, sondern durchaus häufig auftreten, die sogenannten «Halluzinationen». Nicht selten werden Beispiele für diese Tatsache dann im Sinn der in Abschnitt 1 angerissenen schwarz-weißen Betrachtung für die Suggestion verwendet, dass «die Technik das eben einfach nicht kann» und dass die ganze Aufregung es gar nicht wert sei. Nach Ansicht des Verfassers liegt dieser Betrachtungsweise ein grobes Missverständnis zugrunde! Wenn die Erwartung ist, die Maschine mache keine Fehler bzw. sei so gut wie der Mensch – was selbstredend nicht dasselbe ist! –, dann ist die Enttäuschung vorprogrammiert. Wenn die Erwartung hingegen die ist, es hier mit einem Assistenzsystem zu tun zu haben, das die Arbeit in der Regel erleichtert oder beschleunigt oder einfach nur Spaß macht, mit einem System also, dessen Ergebnisse stets sorgfältig durch den Menschen geprüft werden müssen, dann ergibt sich ein anderes Urteil. Enttäuschung ist eine Funktion der Erwartung!

3. Zur Qualität der Ausgaben von ChatGPT

So, wie bei KI im Allgemeinen die wesentliche Frage nicht ist, «ob KI das kann oder nicht», sondern «wie gut KI das in einem bestimmten Kontext kann oder nicht», stellt sich die Situation auch speziell für ChatGPT dar. Und genau wie bei KI-Anwendungen im Allgemeinen gibt es Anwendungsfälle, bei denen ein objektivierter Qualitätsbegriff nicht wirklich wichtig ist: Wenn ein erster Wurf etwa eines Grußworts erstellt ist, kann dieses Grußwort verworfen, angepasst, gemeinsam mit ChatGPT verbessert oder direkt verwendet werden. Wenn ChatGPT Vorschläge für Gute-Nacht-Geschichten unterbreitet hat, können diese übernommen oder verworfen werden. Wenn ChatGPT die Idee für das Drehbuch einer Wissenschaftssendung erstellt hat, kann diese ebenfalls verworfen oder ggf. wiederum gemeinsam mit ChatGPT ergänzt oder angepasst werden. Ein Mensch kann also intuitiv beurteilen, ob die Qualität in einem gegebenen Zusammenhang hinreichend ist oder nicht – und dieses intuitive Verständnis ist häufig vollkommen ausreichend.

Gleichzeitig gibt es viele Situationen, in denen Qualität eine größere Rolle spielt. Arztbriefe, die auch heute oft mithilfe von Textbausteinen erstellt werden, bedürfen natürlich einer sorgfältigen Durchsicht. Klageschriften sollten dem konkreten Fall angemessen sein.[vii] Wenn Programmcode erstellt wird, sollte dieser Code das zu lösende Problem tatsächlich für alle möglichen Eingaben lösen und in diversen anderen Dimensionen[xii] «gut»

oder zumindest adäquat sein.[2] Wenn ChatGPT als Tutorsystem verwendet wird, sollten die Antworten vor allem dann korrekt sein, wenn die lernende Person nicht in der Lage ist, die Richtigkeit oder Plausibilität der Aussagen zu bewerten, was insbesondere für Kinder und Jugendliche der Fall sein dürfte.

Was in einem bestimmten Anwendungsfeld «gut» ist, ist in der Regel stark geprägt durch dieses Anwendungsfeld und vieldimensional häufig mit Konflikten zwischen diesen Dimensionen. Eine Herausforderung stellt die Überlegung dar, wann ChatGPT in einem gegebenen Zusammenhang *hinreichend gut* ist. Das Beispiel einer anderen Art von KI, der Diagnose durch einen Arzt, zeigt das sehr plastisch: ist ein maschinelles Diagnosesystem «gut genug», wenn es so gut ist wie der schlechte Arzt um die Ecke? Wie ein durchschnittlicher Arzt? Oder wie der internationale Spitzenmediziner? Fragen dieser Art sind komplex. Wenn hypothetisch 20% der Aussagen von ChatGPT im Einsatz als Tutorsystem in der fünften Klasse nachweislich falsch sind, darf dieses System dann überhaupt eingesetzt werden? Eine intuitive Antwort mag «natürlich nicht» lauten – aber wenn dem Kind keine anderen Möglichkeiten der Interaktion zur Verfügung stehen, kann man den Einsatz dann nicht rechtfertigen? Oder gibt es eine Schwelle für die Anzahl inkorrekter Antworten, die noch rechtfertigbar ist – im Wissen, dass wir alle unabsichtlich bisweilen Falsches sagen, sei das als Eltern, als Lehrer oder als Universitätsprofessoren? Hier sei auch auf das (problematische) utilitaristische Argument verwiesen, ein autonomes Fahrzeug sei dann hinreichend sicher, wenn es im Schnitt so sicher fahre, sei wie der Mensch. Eine Diskussion über die Notwendigkeit nicht nur utilitaristischer, sondern auch deontologischer ethischer Argumente würde den Rahmen dieses Textes sprengen.

Es ist i.A. sehr schwierig, die Qualität der Ausgaben von ChatGPT oder vergleichbaren Systemen zu objektivieren und zu bewerten. Das liegt nicht zuletzt daran, dass die Fragestellungen, die solche Systeme behandeln, sehr breit gestreut sind und die Problemstellung dementsprechend nicht präzise fassbar ist, was auch die Definition und Bemessung von Qualität a priori erschwert (s. die einführenden Bemerkungen zur Qualität von KI in Abschnitt 1). Ein Ansatz zur Bewertung besteht darin, ChatGPT standardisierte Prüfungen durchführen und ohne Wissen des maschinellen Verfassers durch Menschen bewerten zu lassen. Das Bayerische Abitur 2023 etwa wäre von ChatGPT mit der Note «gut» bestanden worden.[viii] ChatGPT hat auch standardisierte internationale Tests absolviert und liegt nicht selten in den Top-Ten im Vergleich mit Menschen, die denselben Test absolviert haben.[ix] Der IQ von GPT-4 ist 124.[x]

Zusammenfassend ist festzuhalten, dass ChatGPT konstruktionsgemäß nicht immer korrekte oder adäquate Antworten liefern kann. Nach Ansicht des Verfassers mündet das in die wichtige gesellschaftliche Fragestellung, welche Qualität wir in welchem Kontext denn für hinreichend halten. Oft erkennen wir das intuitiv sofort, auch ohne Objektivierung dessen, was gut genug ist. In jedem Fall sind ChatGPT und vergleichbare Systeme als Assistenten zu begreifen, deren Ausgabe immer durch einen Menschen überprüft werden muss – und die Qualität der Ausgabe hängt von der Qualität des Prompts ab. Der Einzelnachweis, dass ChatGPT eine bestimmte Frage falsch beantwortet, ist wohlfeil und in der Debatte nicht zielführend.

4. Die Technik hinter ChatGPT

Nach Ansicht des Verfassers war Ende 2022 ein gewisses Abflachen des KI-Hypes zumindest im deutschsprachigen Raum beobachtbar. Mit der Veröffentlichung von ChatGPT, in den Worten des Vorstandsvorsitzenden von NVIDIA dem «iPhone-Moment der KI», wurde ein neuer Zyklus des Hypes gestartet. Künstliche Intelligenz, eine nach Ansicht des Verfassers geniale Bezeichnung, die gleichzeitig der menschlichen Intelligenz sicher nicht

[2] Die Qualität der Erzeugung von Code durch KI wird häufig über Benchmarks gemessen: Eine vorgegebene Anzahl von Programmieraufgaben wird mit KI gelöst, und dann wird über Tests näherungsweise herausgefunden, ob der Code das tut, was er tun soll. Die Anzahl der richtigen Lösungen befindet sich zwischen 20 und 80 Prozent. Benchmarks u.a. bei https://arxiv.org/pdf/2202.13169.pdf und https://arxiv.org/abs/2107.03374, s. auch https://dl.acm.org/doi/10.1145/3520312.3534864 zur Produktivität.

gerecht wird, ist eine spezielle Form von Software, also Computerprogrammen. In klassischen Programmen wird die von einem Menschen erdachte Lösung eines Problems in immer kleinere Schritte zerlegt, die ebenfalls durch einen Menschen sehr präzise formuliert werden müssen. Mit maschinengelernten Modellen, der ChatGPT zugrundeliegenden KI-Technologie,[xi] wird hingegen der Ansatz verfolgt, nicht den Lösungsweg zu formulieren, sondern diesen anhand von Beispielen zu lernen und existierende Beispiele, die Trainingsdaten, möglichst adäquat zu generalisieren.[xii]

ChatGPT als ein Vertreter der sogenannten generativen KI basiert auf einem speziellen Verfahren des Maschinenlernens, sog. neuronalen Netzen, hier speziell großen Sprachmodellen (Large Language Models, LLMs).[xiii] ChatGPT besteht aus zwei Teilen, einer Schnittstelle für die «Unterhaltung» (Chat) mit dem neuronalen Netz, und dem LLM, einem sog. «Generative Pretrained Transformer» (GPT).[3] Neuronale Netze bilden Eingaben auf Ausgaben ab, berechnen also mathematische Funktionen. Das können Klassifizierungen von Objekten sein (Bilder oder Videos werden auf die Klassen Fußgänger, Auto, Straßenschild, Entscheidung «defektes Teil» oder «nicht defekt» abgebildet), Positionsangaben (Bilder oder Videos werden auf die Positionen von Fußgängern, Autos, Straßenschildern abgebildet), Vorhersagen über die Leistung von Studierenden (Lebensläufe werden auf Noten im Abschlusszeugnis abgebildet), schwierige Optimierungsprobleme (Landkarten und Anfangspositionen werden auf möglichst kurze Wege abgebildet) usw. Im Fall von LLMs werden Texte, Bilder, Videos, Audio, Code auf Texte, Bilder, Videos, Audio, Code abgebildet. So bildet ChatGPT den oben eingeführten Prompt, die textuelle Eingabe, auf einen (Antwort-)Text ab; Midjourney (und inzwischen auch der ChatGPT-Assistent in Microsofts Suchmaschine Bing) bildet Texte auf synthetische Bilder ab; CoPilot bildet Codefragmente u.a. auf mögliche Fortsetzungen des Codes oder Erklärungen des Codes ab.[xiv]

Vor ihrer Verwendung müssen neuronale Netze trainiert werden. Technisch besteht ein neuronales Netz grob gesagt aus Neuronen und Verbindungen zwischen diesen Neuronen, die in hintereinanderliegenden Schichten angeordnet werden. Es gibt unterschiedliche Architekturen solcher Netze. Im Wesentlichen werden Neuronen einer Schicht mit Neuronen der direkt dahinterliegenden Schicht verbunden, so dass jedes Neuron eine Menge von Verbindungen zu den Neuronen der davorliegenden und der dahinterliegenden Schicht aufweist. Rückkopplungen sind in manchen Architekturen neuronaler Netze ebenfalls üblich. Im Fall von ChatGPT in der Version 3.5 gibt es 96 solcher Schichten und ca. 175 Milliarden Verbindungen zwischen den Neuronen.

Die Verbindungen sind mit einer Verbindungsstärke versehen, einer Zahl, die während des sogenannten Trainings berechnet wird. Im Training wird für die Trainingsdaten der Fehler der Vorhersage durch das Netz minimiert. Wie genau dieser Prozess aussieht, kann hier nicht ausgeführt werden. Trainiert werden die LLMs jedenfalls mit großen Datenmengen, im Fall von GPT 3.5 mit etwa 600 Gigabyte Text vornehmlich aus dem Internet, was der Textmenge von ungefähr 120.000 Bibeln entspricht. GPT wird in einem zweiten Schritt manuell trainiert, dem sogenannten überwachten Lernen, in dem die Reaktion des LLM auf vorgegebene Prompts von Menschen händisch überprüft wird. Im Wesentlichen wird die positive oder negative Rückkopplung durch den Menschen verwendet, um die Verbindungsstärken zwischen den Neuronen zu verändern, so dass bei Vorliegen einer negativen Rückkopplung zukünftig eine andere Antwort gegeben wird. In einem dritten wiederum automatisierten Schritt findet sogenanntes Reinforcement Learning statt, in dem das Modell aus Reaktionen auf die Ausgabe selbst feststellt, wie seine vorherige Antwort nicht noch verbessert werden kann und diese Information wiederum für eine Modifikation der Verbindungsstärken verwendet.[xv]

Die erste Schicht des neuronalen Netzes ist die Eingabeschicht, an die die Eingaben als eine Liste von Zahlen angelegt werden, für jedes Neuron der Eingangsschicht eine solche

[3] Zur Abgrenzung: Neuronale Netze sind eine populäre Technik der KI. LLMs sind spezielle neuronale Netze, die im Rahmen der generativen KI eingesetzt werden, deren Ziel das Erzeugen von Text, Audio, Video, Code ist. GPT ist ein LLM. ChatGPT ist GPT in den Versionen 3.5 oder 4 mit einer natürlichsprachlichen Schnittstelle. Aus sprachlichen Gründen differenzieren wir nicht immer sauber zwischen den Konzepten.

Zahl. Im Fall von ChatGPT repräsentiert diese Liste von Zahlen die Wörter des Prompts, die auf eine spezielle Art und Weise für jeden Prompt in eine solche Liste von Eingabezahlen umgewandelt werden. Die Ausgabeschicht codiert das Ergebnis der Berechnung. Im Fall von ChatGPT repräsentiert jedes Ausgabeneuron genau ein Wort, für das eine Wahrscheinlichkeit berechnet wird, die im Wesentlichen darstellt, was bezüglich des «Wissens» von ChatGPT das wahrscheinlichste, das zweitwahrscheinlichste, drittwahrscheinlichste usw. nächste Wort des Prompts ist. Eins der wahrscheinlichsten Fortsetzungswörter wird ausgewählt, hinten an den Prompt gestellt, und der resultierende neue Prompt wird wieder als Eingabe ausgewählt und das wahrscheinliche zweite Folgewort berechnet. Dieses wird wiederum an den ergänzten Prompt angehängt, um das dritte wahrscheinlichste Wort zu berechnen usw. Wenn nicht stets das nächste wahrscheinlichste Wort ausgewählt wird, sondern ein etwas weniger wahrscheinliches, kann so die wahrgenommene «Originalität» des Texts erhöht werden.

Aus dieser grob vereinfachten Darstellung sieht man sofort, dass GPT über kein «Verständnis» in einem intuitiven Sinn verfügt, sondern «nur» wahrscheinliche nächste Wörter berechnet. Die Technologie wird entsprechend auch als «stochastischer Papagei» bezeichnet. Dass ChatGPT häufig Ausgaben von verblüffender Qualität erfolgt, erstaunt auch diejenigen, die die Technik vollständig verstehen. Die Tatsache, dass ChatGPT immer das nächste wahrscheinlichste Wort berechnet wird, erklärt auch das Phänomen des o.g. Halluzinierens, die Berechnung von in der Regel eloquent formulierten, aber vollständig falschen oder gar sinnbefreiten Antworten. Schließlich hängt die Qualität der Ausgabe maßgeblich, wie bereits erwähnt, von Inhalt und Form des Prompts ab.

Aus der Technologie ergibt sich auch, dass ein tiefergehendes mathematisches Verständnis nicht zu erwarten ist. Der Trend geht deswegen heute in vielen Anwendungsgebieten dahin, ChatGPT mit anderen Werkzeugen zu koppeln – mathematischer Software, Internetrecherche, Werkzeugen der Softwareentwicklung.

5. Auswirkungen von ChatGPT

Die Diskussion um die gesellschaftlichen Auswirkungen von ChatGPT ist wie die Diskussion um KI allgemein bisweilen von Hysterie und Polarisierung geprägt. Wie die Technologie der LLMs die Gesellschaft verändern wird, kann niemand vorhersagen. Erste Einschätzungen und informierte Spekulationen scheinen mittlerweile aber möglich. Wie viele andere ist auch der Verfasser dieses Textes der Ansicht, dass ChatGPT signifikanten Einfluss auf viele Bereiche unseres Lebens haben wird.

In Abschnitt 2 haben wir Beispiele konkreter Anwendungsfälle skizziert und in Abschnitt 3 ausgeführt, wie kontextabhängig und schwierig es ist, die Qualität von Antworten zu definieren und zu bemessen. Wir haben auch gesehen, dass die Frage nach der hinreichenden Qualität mindestens ebenso schwierig ist. Wir haben dargelegt, dass ChatGPT als Assistenzsystem zu sehen ist, das Menschen kreativer, freudiger und effizienter arbeiten lässt. Die Frage, ob ChatGPT Menschen *ersetzt*, stellt sich nach Ansicht des Verfassers nur in sehr wenigen Fällen, vielleicht bei Katalog-Models oder Schauspielern.[xvi] Häufig zitierte Anwendungsfelder wie etwa die Buchhaltung oder das Verfassen von Emails, Newslettern und Social Media-Posts werden nach Ansicht des Verfassers immer von Menschen begleitet werden, die Plausibilität, Korrektheit, Vollständigkeit und Tonalität überprüfen werden (müssen). Hier gilt wohl der mittlerweile etwas abgedroschene Slogan, dass ein Mensch nicht durch generative KI ersetzt wird, sondern durch einen anderen Menschen, der generative KI verwendet.

Wir haben bereits argumentiert, dass wohl in vielen Anwendungsfeldern davon auszugehen ist, dass eine Verschiebung vom Schaffenden hin zum Überprüfenden erfolgen wird, was insbesondere die Notwendigkeit zeitigt, über die Fähigkeit zum kritischen Reflektieren, zum Überprüfens und die Urteilskraft nachzudenken.[xvii] Prompt Engineering allein wird in den meisten Fällen sicherlich nicht ausreichen, auch wenn natürlich das iterative Verbessern eines Prompts in der Regel durch Unzufriedenheit mit dem Ergebnis veranlasst sein wird.

Es gibt unzählige Anwendungsfelder, und jedes einzelne Anwendungsfeld verdient eine umfassende Betrachtung, der wir hier nicht gerecht werden können. Im Rahmen dieses Texts wollen wir deswegen für einige wenige ausgewählte Anwendungsfelder relevante Perspektiven skizzieren, ohne den Anspruch auf Vollständigkeit zu erheben. Es wird auch klar werden, dass für viele Herausforderungen heute noch keine überzeugenden Antworten existieren.

5.1. Lehren und Lernen

Ein wichtiges Anwendungsfeld ist das Lehren und Lernen, das wir in der Diskussion um die Qualität von ChatGPT als Tutorsystem bereits angerissen haben. Um nur zwei Beispiele zu nennen, ist ChatGPT nach Ansicht des Verfassers erstaunlich gut darin, Code zu erklären, aber auch darin, etwa grammatikalische Analysen lateinischer Sätze durchzuführen. Natürlich können die Aussagen von ChatGPT falsch sein, was in einem Lehrer-Schüler-Verhältnis herausfordernd ist, weil die Schüler ja eigentlich der fachlichen Autorität des Lehrers vertrauen können sollten. Gleichzeitig eröffnet die Möglichkeit des Dialogs mit ChatGPT enorme didaktische Spielräume, vor allem dann, wenn dialogisches Lernen mit ChatGPT zu jeder Tageszeit und in beliebigem Tempo mit dem heute häufig vorherrschenden Frontalunterricht kontrastiert wird. Nach Ansicht des Verfassers ist davon auszugehen, dass grundlegende Zusammenhänge in (Berufs-)Schule und Studium in naher Zukunft mit guter Qualität durch ChatGPT auch im Dialog gelehrt werden können – was heute noch nicht der Fall ist, wenn etwa das Werkzeug Perplexity fälschlich erklärt, dass «Längengrade vom Nord- zum Südpol verliefen, *also von links nach rechts*».

Unabhängig vom gelehrten bzw. gelernten Fach stellt sich die Frage, welche Fähigkeiten Lernenden beigebracht werden müssen. Wir haben gesehen, dass die Fähigkeit zum Überprüfen von Antworten Teil einer allgemeineren Urteilsfähigkeit ist, die entsprechend trainiert werden muss – sofern das überhaupt möglich ist und eine solche Urteilsfähigkeit sich nicht ohnehin erst in der Praxis einstellen kann (und dass Urteilskraft auch Herzensbildung ist, vernachlässigen wir hier). Grob gesagt kann nur der- oder diejenige Programmcode bewerten, der oder die selbst programmieren kann. Wir gehen davon aus, dass sich das in anderen Fachgebieten ganz ähnlich verhält. Für die Ausbildung bedeutet dies, dass wir bzgl. der zu lernenden Inhalte vielleicht gar nicht so viel verändern müssen. Die Fähigkeit zur kritischen Analyse hingegen kann vielleicht im Sinn einer Kompetenzorientierung noch stärker gefördert werden, also im Fall von Programmcode die Fähigkeit, den Code als funktional korrekt, adäquat und performant anzusehen, was sich in Techniken wie dem Lesen oder Testen von Code niederschlägt.

5.2. Prüfungswesen

Eng verbunden mit neuen Möglichkeiten und Stolpersteinen in der Ausbildung ist das Abprüfen des Gelernten. Wir haben bereits gesehen, wie gut ChatGPT standardisierte Prüfungen oder das Abitur bestehen kann. Auch Klausuren zu Informatikvorlesungen werden von ChatGPT mit «gut» bestanden. Solange sichergestellt werden kann, dass Studierende Prüfungen ohne Hilfe von ChatGPT anfertigen, ändert sich scheinbar nicht viel. Allerdings ist die Notwendigkeit von Fernprüfungen – und übrigens auch die damit einhergehende organisatorische Erleichterung – während der Corona-Pandemie offenbar geworden; entsprechende Gesetze wurden verabschiedet. Wir haben bereits bemerkt, dass derzeit nicht mit hinreichender Genauigkeit automatisiert entschieden werden kann, ob ein Text menschlichen oder maschinellen Ursprungs ist. Es stellt sich also die Frage, wie in Zeiten von ChatGPT Prüfungen oder Teile davon abgelegt werden sollen, wenn nicht sichergestellt werden kann, dass ChatGPT für die Bearbeitung verwendet wurde.

Nicht nur Klausuren im Studium, auch Hausaufgaben von der Primarschule bis zur Universität sind eine Form der Prüfungsleistung. Wenn die Versuchung zu groß ist, diese Leistungen nicht selbst, sondern durch oder mit ChatGPT erstellen zu lassen, sind gutgemeinte Ratschläge der Art, dass man ja nun einmal für das Leben und nicht die Schule lerne, eingeschränkt hilfreich. Wir haben an anderer Stelle die verschiedenen Aspekte und Zielsetzungen von Prüfungen im Zusammenhang mit ChatGPT dargelegt.[xviii] Mündliche Leistungskontrollen sind prinzipiell geeignet, wirkliches Verständnis abzufragen. Sie

skalieren aber nicht, was im Fall von Anfängervorlesungen in Informatik an der TU München deutlich wird, die regelmäßig mit mehr als 2000 Prüflingen stattfinden.

Der Versuch der Durchsetzung eines Verbots von ChatGPT für Prüfungsleistungen scheint uns in der Regel aus praktischen Gründen letztlich auch perspektivisch zum Scheitern verurteilt. Alternativ kann der Einsatz von ChatGPT in der Prüfung explizit verlangt werden und die Reflexion über den Dialog mit dem Werkzeug Teil der Prüfungsleistung sein. So wird der Einsatz des Systems nicht nur Teil der Lehre, sondern die kritische Reflexion Teil der Prüfung.

Es wird nicht allen leichtfallen, der Versuchung zu widerstehen, Haus-, Seminar- und Abschlussarbeiten mit Hilfe von ChatGPT zu erstellen. Abhängig vom Fach kann man sich aber fragen, wie schlimm das eigentlich ist. Wenn Prüfende mit Texten konfrontiert werden, die dank ChatGPT von akzeptabler Qualität sind, was nicht immer selbstverständlich ist, dann gibt es mindestens zwei Perspektiven, aus denen ChatGPT als hilfreich anerkannt werden kann. Das hängt natürlich vom Fach ab. In der Informatik etwa ist der wesentliche Beitrag einer Abschlussarbeit häufig ein Computerprogramm, ein Experiment, ein Konzept oder eine empirische Untersuchung. Der Text, der diesen Beitrag der Abschlussarbeit beschreibt, ist dann nur ein Teil der Prüfungsleistung, was in anderen Fachgebieten anders sein mag, wo der Text selbst die zentrale Leistung darstellt. Gleichwohl ist unbenommen, dass die Fähigkeit zur Destillierung des Wesentlichen und dessen sprachliche Ausgestaltung natürlich eine wesentliche Kompetenz darstellen und trainiert werden müssen.

Schließlich kann ChatGPT umgekehrt gewinnbringend auch für die Formulierung und Vorabdiskussion von Prüfungsfragen verwendet werden.

5.3. Dokumentation

Attraktiv erscheint die Verwendung von ChatGPT zur Erstellung von ausformulierten Dokumentationen, die in vielen Lebensbereichen gefordert und/oder sinnvoll sind. Die grundlegende Idee ist, dass im Prompt die wesentlichen Informationen vorgegeben und die durch ChatGPT erstellte Dokumentation durch einen Menschen überprüft und unterschrieben wird. Sinn der Dokumentation ist ja letztlich die Möglichkeit der Zuschreibung von Verantwortung an Menschen. Auf den ersten Blick mag der Ansatz unsinnig erscheinen, weil es ja zunächst scheinbar ausreichen würde, allein den Prompt abzuspeichern, aus dem die Dokumentation beliebig oft erzeugt werden kann. Prompts sind aber üblicherweise im Wesentlichen Stichworte, die viel Interpretationsspielraum lassen – und die Ausformulierung engt diese Spielräume ein, wie jeder weiß, der selbst einen Text auf der Basis von Stichpunkten ausformuliert. Entscheidend ist in jedem Fall die abschließende Prüfung durch den Menschen und seine Unterschrift.

5.4. Urheberschaft und -recht

Wenn ChatGPT einen Text erzeugt, wer ist dann im rechtlichen Sinn dessen Urheber? Im deutschen Recht kann das nur ein Mensch sein.[xix] In Abschnitt 4 haben wir gesehen, dass ChatGPT Prompts im Wesentlichen jeweils mit einem nächstwahrscheinlichen Wort fortsetzt, was die Frage nach einer wie auch immer gearteten schöpferischen Leistung aufwirft. Wir wollen uns hier der Diskussion entziehen, ob oder in welchem Sinn ChatGPT kreativ ist. Gleichwohl stellt sich die Frage, was die Verwendung eines Werkzeugs wie ChatGPT fundamental von der Verwendung etwa einer Internetrecherche oder das begleitende Lesen von Büchern bei der Erstellung eines Textes unterscheidet. Auch hier werden letztlich Ideen aus existierenden Quellen aufgegriffen, mit anderen Ideen amalgamiert und in eine Argumentation gegossen. Natürlich ist die Auswahl der Quellen und die Destillierung des Wesentlichen darin als solches ein schöpferischer Prozess – aber macht es wirklich einen fundamentalen Unterschied, ob dieser Prozess durch einen Menschen oder eine Maschine durchgeführt wird, und sei dieser Prozess nur statistischer Natur?

Wie im Feld der Medien herrscht heute jedenfalls im Bereich des Software Engineering Skepsis, ob durch generative KI erzeugter Code überhaupt verwendet wird und das nicht ggf. doch urheberrechtliche Konsequenzen hat. Viele Unternehmen trauen sich schlicht nicht, solchen generierten Code zu verwenden. Microsoft, das hinter dem Werkzeug

CoPilot steht, übernimmt deswegen derzeit prinzipiell die Verantwortung für zukünftige gerichtliche urheberrechtliche Auseinandersetzungen.[xx]

Vielleicht ist es aber auch angezeigt, den Begriff des Urheberrechts generell zu überdenken, was offensichtlich notwendig und nicht nur der Tatsache geschuldet ist, dass zumindest im deutschen Recht Maschinen als Urheber nicht vorgesehen sind. Wir haben an anderer Stelle[xvii] darauf hingewiesen, dass der Begriff der Urheberschaft kulturellen Prägungen unterliegt und auch durchaus widersprüchlich ist. Alten Meistern etwa werden Kunstwerke wie Statuen oder Bilder allein zugeschrieben, auch wenn natürlich ein ganzer Stab an Mitarbeitern maßgeblichen Anteil an diesen Kunstwerken hatte. Reden und Bücher werden häufig von Ghostwritern verfasst, Assistenten an der Hochschule helfen bei der Vorbereitung von Vorlesungsfolien. Und jeder Mensch macht sich bewusst oder unbewusst Ideen zu eigen, wenn er Zeitung liest, Diskussionen führt oder sich Texte wie diesen einverleibt – häufig, ohne die Quelle dieser Ideen anzugeben und bisweilen nach einer gewissen Zeit in der Überzeugung, diese Idee selbst gehabt zu haben.

Wichtigen Raum in dieser Diskussion nimmt die Frage nach der Kennzeichnung von Quellen ein, was etwa in der Wissenschaft konstitutiv ist. Es ist offenkundig, dass ein solcher Verweis auf Quellen angesichts der statistischen Natur der Erzeugung von Ausgaben prinzipiell schwierig, wenn nicht unmöglich ist. Ein populäres Vorgehen heute ist vereinfacht gesagt, die von ChatGPT erzeugte Ausgabe als Anfrage an eine Internet-Suchmaschine zu schicken und die relevantesten Resultate als Quellen anzugeben. Ob das dem Wunsch oder der Pflicht gerecht wird, Quellen angemessen zu zitieren, sei dahingestellt.

5.5. Graphisches und textuelles Design

In der regelmäßigen Erstellung von Texten wie etwa Newslettern, Emails oder Social Media Posts ist die Verwendung von ChatGPT heute anekdotisch bereits weitgehend der Normalfall – auch wenn es einzelne Fälle gibt, bei denen die «Manufaktur» von Texten ohne Verwendung von KI als Verkaufsargument verwendet und zur Rechtfertigung höherer Preise in Stellung gebracht wird. In welchen Fällen Kunden willens sind, den höheren Preis für solche rein händisch – oder auch hybrid von Mensch und Maschine – erzeugten Produkte zu bezahlen, wird sich herausstellen. Anforderungen an Qualität und Angemessenheit der Tonalität der Texte sind sicherlich von den konkreten Gegebenheiten abhängig. ChatGPT wird auch verwendet, um Bücher, Drehbücher,[xvi] Geschichten oder Gedichte zu verfassen – oder längere Texte zusammenzufassen,[vi] was möglicherweise rechtliche Herausforderungen birgt. Ob auch hier die Qualität jeweils hinreichend gut ist, hängt natürlich wiederum vom Einsatzgebiet ab.

Erneut sei kurz darauf verwiesen, dass LLMs auch zur Synthese von Bildern oder Videos verwendet werden können. Typische Beispiele sind die Erstellung von Bildern nicht nur in der Werbung, von multilingualen Firmenvideos[xxi], oder von vollständigen Videos aus ebenfalls von LLMs erstellten Drehbüchern. Wir sehen enormes Potential in der entsprechenden Technologie, müssen hier aber aus Platzgründen auf eine weitere Erörterung verzichten.

5.6. Software Engineering

Software Engineering beinhaltet neben dem Erstellen von Code viele andere Aktivitäten etwa im Requirements Engineering, dem Design von Architekturen, der Erklärung, Optimierung und Umstrukturierung von Code, dem Testen, der Dokumentation und der Wartung. Spezielle für die Erstellung von Code geschaffene LLMs werden schon seit geraumer Zeit in unterschiedlichen entsprechenden Anwendungen untersucht.[xxii] Generell gilt offensichtlich, dass LLMs die gestellten Aufgaben im Software Engineering dann zufriedenstellender lösen können, wenn dieselbe oder ähnliche Aufgaben in der Vergangenheit bereits gelöst wurden und dieses Wissen gleichsam beim Training in das LLM «eingebacken» wurde. Dies ist übrigens erstaunlich häufig der Fall.

Bei der Codeerzeugung erstellen LLMs Code aus einer Aufgabenstellung heraus oder unterbreiten Vorschläge für die Fortsetzung bereits geschriebenen Codes. Im Unterschied zu anderen Anwendungsgebieten von LLMs ist eine besondere Herausforderung hier, dass der Code exakt das tun muss, was die Aufgabenstellung verlangt, also für jede Programmeingabe fehlerfrei funktionieren muss. Der Bewertung des erzeugten Codes durch Lesen

oder durch Testen kommt also besondere Wichtigkeit zu. In dieser Hinsicht ist es fast erstaunlich, wie gut die Erstellung von Code mit LLMs funktioniert! Der Hersteller des Werkzeugs CoPilot etwa berichtet, dass viele Code-Vorschläge angenommen werden und dass, wichtiger vielleicht, Software-Ingenieure ihre Arbeit mit CoPilot als durchgängiger und produktiver empfinden.[xxiii] Anekdotische Evidenz suggeriert, dass erzeugter Beispielcode sehr nützlich ist, um die Funktionsweise neuer Technologien zu verstehen; dass es manchmal aber einfach schneller ist, den Code selbst zu schreiben, anstatt iterativ einen Prompt zu verbessern und jeweils das Ergebnis überprüfen zu müssen. Es gibt den Scherz, dass früher 23 Stunden codiert und eine Stunde getestet wurde – und heute eben mit LLMs eine Stunde codiert und 23 Stunden getestet wird. Die Qualität des erzeugten Codes hängt natürlich wiederum von der Detailtiefe und Strukturierung des Prompts ab. Eine Aufgabe von Programmierern ist es, Konzepte aus dem fachlichen Anwendungsbereich der zukünftigen Software in technische Konzepte zu übersetzen. Nach Ansicht des Verfassers wird es hier noch einige Zeit dauern, bis dieser kreative Schritt von Maschinen übernommen werden kann – wenn das überhaupt jemals der Fall sein kann.

LLMs scheinen auch sehr nützlich bei der Erzeugung von schablonenhaftem Code etwa für Tests zu sein. Bei der Testlogik und den verwendeten Eingabewerten für Tests hingegen ist es ebenfalls nicht immer klar, ob die nicht schneller durch einen Menschen erzeugt werden. Anekdotisch wird weiterhin großes Potential in der Erzeugung der Dokumentation sowie in der Erklärung von Code gesehen.

Festzuhalten ist, dass die Verwendung von LLMs in die tägliche Praxis der Software-Entwicklung bereits Einzug gehalten hat. Wie weit das Automatisierungspotential reicht, wird sich zeigen. In jedem Fall ist wohl davon auszugehen, dass Assistenten wie CoPilot zukünftig ein selbstverständlicher Teil der Entwicklungsumgebungen von Software-Ingenieuren sein werden. Gleichzeitig werden Software-Ingenieure sicherlich nicht durch LLMs ersetzt werden: Ein großer Teil des Software-Engineering besteht in der Kommunikation mit Auftraggebern, dem Verständnis von Anforderungen, der Festlegung adäquater Strukturen, die Übersetzung von Konzepten aus dem Wirkbereich der Software in technische Strukturen und der Überprüfung, ob ein Softwaresystem das tut, was es tun soll – was alles Aktivitäten sind, für die wesentliche Unterstützung durch LLMs nur schwierig vorstellbar ist.

6. Risiken

Dieser Aufsatz nimmt bewusst und aus tiefer Überzeugung eine eher chancenorientierte Perspektive auf ChatGPT und verwandte Technologien ein. Zum Abschluss wollen wir dennoch auf einige wohlbekannte Risiken hinweisen.

Auf die gesellschaftlichen Auswirkungen von Deep Fakes mit entsprechenden computergenerierten Texten haben wir bereits hingewiesen. Wenn es einfach möglich ist, etwa politische Konkurrenten mit fragwürdigen Situationen oder Aussagen in Verbindung zu setzen, wird es angesichts der Wirkmacht von Bildern und Videos schwierig sein, einen einmal erzeugten Eindruck zu korrigieren. Vielleicht wird es technisch möglich sein, Deep Fakes mit hinreichend guter Genauigkeit zu erkennen oder zumindest Nutzer auf eine erhöhte Wahrscheinlichkeit hinzuweisen, dass ein Video ein Fake ist. In jedem Fall gilt auch hier bestimmt, dass ein ausgebildetes Urteilsvermögen hilfreich ist.

Häufig wird darauf hingewiesen, dass durch die Konstruktion von LLMs eine Mehrheitsmeinung zementiert wird. ChatGPT etwa konnte eine linksliberale «Meinung» zugeschrieben werden,[xxiv] die darin gründet, dass dies die im Internet verbreitetere politische Ausrichtung zu sein scheint. Es gibt auch die Sorge, dass die Kreativität der Menschheit eingeschränkt wird, wenn im Wesentlichen Texte erstellt werden, die auf bereits bestehenden Texten basieren. Der Verfasser ist nicht sicher, wie groß dieses Problem wirklich ist, wenn nicht zuvor herausgefunden wird, wieviel von Menschen erstellter Text in der Vergangenheit ebenfalls letztlich bestehende Ansichten reproduziert hat.

Eine weitere Sorge betrifft den sogenannten Automation Bias. Es ist bekannt, dass Menschen nachlässig oder unaufmerksam werden, wenn von einer Maschine gemachte Vorschläge fast immer richtig oder adäquat oder hinreichend gut sind. Die simple

Aufforderung zu geistiger Regsamkeit hilft da natürlich nicht. Eine Fortsetzung dieses Belangs ist das sog. Deskilling, das die Tatsache bezeichnet, dass Menschen bestimmte Fähigkeiten verlieren, wenn sie durch Maschinen erledigt werden können. Das Standardbeispiel ist das Lesen von Landkarten, wenn es Navigationssysteme gibt. Der Verfasser ist nicht sicher, ob das wirklich ein großes Problem darstellt und verweist darauf, dass bereits in der Antike die Erfindung der Schrift kritisiert wurde, da damit die Menschen ihr Erinnerungsvermögen verlören.

Schließlich gibt es die Sorge, dass mit LLMs Schädlinge geschaffen werden können – natürliche Schädlinge im Sinn etwa von «Killerviren» und künstliche Schädlinge im Sinn von Schadsoftware. Zu ersteren kann der Verfasser keine Aussage treffen. Zu zweiteren steht der Nachweis noch aus, dass schwer zu erkennende und großen Schaden anrichtende Malware durch LLMs entwickelt werden kann. Was ein Problem darstellen mag, ist die Verfügbarkeit von LLMs und somit die noch einfachere Erstellung von Malware durch jedermann.

ChatGPT eröffnet zahlreiche ethische Fragestellungen. Wir haben bereits die Frage angesprochen, ob ChatGPT als Tutorsystem verwendet werden darf, wenn falsche Aussagen nicht nur die Ausnahme sind. Es ergeben sich offensichtliche Fragestellungen – eher einfache der Art «wie soll eine mit ChatGPT ausgestattete Puppe auf Fragen nach dem Weihnachtsmann reagieren?»; komplexere wie «soll ChatGPT als Dialogpartner für Demenzkranke fungieren dürfen?»; und noch anspruchsvollere der Art «sollen wir bösartige Schaftsoftware erzeugen lassen, um zu lernen, wie wir uns verteidigen können?» An dieser Stelle überlassen wir die Fragen Kundigeren.

Wenngleich vielleicht kein Risiko im engeren Sinn, betrifft die größte Sorge vermutlich die Unsicherheit, welchen Einfluss LLMs auf die Perspektive bestimmter Berufsgruppen hat. Wir haben im Text wiederkehrend darauf hingewiesen, dass einzelne Gruppen möglicherweise in der Tat ersetzt werden können; dass aber in den meisten Fällen davon auszugehen ist, dass LLMs mächtige Assistenten sind, die aber eben in dieser Funktion der menschlichen Steuerung und Überprüfung bedürfen.

Zuletzt sei die Risikoeindämmung durch gesetzliche Regulierung angesprochen. LLMs haben als sogenannte Foundational Models im Nachhinein Eingang in den derzeit diskutierten AI Act der EU gefunden, sind also ein Werkzeug der KI und werden in der Entwicklung anwendungsbezogen bzgl. ihrer Risikoklasse reguliert. Der risikoklassenbasierte Ansatz ist sicherlich einleuchtend; ob allerdings die Regulierung von Technik selbst anstelle einer sektoralen Regulierung der Anwendung nicht nur politisch gangbar, sondern auch inhaltlich angemessen ist, wird sich herausstellen müssen.

7. Schluss

Generative KI mit LLMs wie ChatGPT erzeugen Texte, Bilder, Videos und Audio in häufig wirklich verblüffender Qualität. Diese Qualität hängt zum einen von der Beschaffenheit der Eingabe, des Prompts ab. Wenn ChatGPT zunehmend als natürlichsprachliche Schnittstelle zu anderen, komplexeren Systemen verwendet wird, müssen wir lernen, entsprechende Prompts zu entwerfen. Zum anderen haben wir gesehen, dass das, was Qualität auszeichnet, ob Qualität relevant ist und was hinreichende Qualität ist, vom Anwendungsgebiet abhängt. Die Chancen, die sich aus dieser Technologie ergeben, scheinen enorm.

Gleichwohl werden LLMs, wie vorher andere Techniken der KI, derzeit gehyped. Die mediale Aufmerksamkeit changiert zwischen «kann alles und muss eingehegt werden» bis zu «letztlich lebensbedrohend». Beides ist übertrieben. ChatGPT ist ein Werkzeug wie andere, ein mächtiges Werkzeug. Viele, den Verfasser eingeschlossen, sind der Ansicht, dass ChatGPT und verwandte Technologien viele Lebensbereiche massiv beeinflussen werden. Wie genau dieser Einfluss aussieht und wo er endet, kann heute niemand sagen – zumal viele technische Schwächen in neueren Sprachmodellen behoben werden.

Assistenzsysteme wie ChatGPT werden diverse Tätigkeiten vom Schaffen hin zum Überprüfen verschieben. Nicht immer ist ausgemacht, dass das entsprechende Vorteile in Effektivität oder Effizienz bringt; das lernen wir gerade durch Ausprobieren. Es ist aber

offenkundig, dass durch ChatGPT viele Aktivitäten enorm beschleunigt oder vereinfacht werden, vor allem dann, wenn ein objektivierter Qualitätsbegriff schwierig zu fassen oder gar nicht notwendig ist. In jedem Fall ergibt sich für uns, dass wir das kritische Reflektieren noch stärker in den Vordergrund stellen müssen. In der Ausbildung ist denkbar, dass kritische Reflexion ohne Beherrschung des zugrundeliegenden Handwerks nicht denkbar ist, wie wir im Fall der Ausbildung von Informatikern argumentiert haben.

Interessant erscheint uns die Debatte um ChatGPT, weil sie sich wie im Fall von Blockchains oder allgemeiner KI um Technik dreht, um eine Infrastruktur also, und nicht um Anwendungen – so wie sich häufig die Debatte um Digitalisierung um das abstrakte Phänomen dreht und nicht konkrete Anwendungen, konkrete Vorteile und konkrete Nachteile. Vielleicht erklärt diese Diskussion im Technisch-Abstrakten auch die bemerkenswerte Polarisierung der Debatte, weil gewissermaßen gleichzeitig über alle Anwendungen gesprochen wird.

Was genau generative KI im Einzelfall können wird und wie das unser Leben verändert, vermag niemand heute zu sagen. Es liegt an uns selbst, uns ein Bild zu machen, die Technologie zu verwenden und ihren Einsatz zu reflektieren.

---

[i] https://chat.openai.com/auth/login

[ii] Beispielsweise https://www.cnbc.com/2018/03/13/elon-musk-at-sxsw-a-i-is-more-dangerous-than-nuclear-weapons.html und https://www.nytimes.com/2023/07/25/opinion/karp-palantir-artificial-intelligence.html.

[iii] https://www.kaggle.com/c/deepfake-detection-challenge

[iv] Einführungen in das Prompt Engineering finden sich etwa bei https://learnprompting.org/, https://www.neat-prompts.com/, https://flowgpt.com/ und https://snackprompt.com/.

[v] In Singapur wurden 90‘000 Bedienstete des öffentlichen Dienst in der Anwendung von ChatGPT ausgebildet, https://mothership.sg/2023/02/singapore-civil-service-chatgpt/; in Baden-Württemberg wird die Textassistenz F13 erprobt, https://stm.baden-wuerttemberg.de/de/service/presse/meldung/pid/kuenstliche-intelligenz-in-der-verwaltung.

[vi] S. etwa https://www.scientific-economics.com/chatgpt-zusammenfassung-schreiben-lassen/. Der Verfasser dieses Textes ist von der Qualität der von ChatGPT erstellten Zusammenfassungen nicht immer überzeugt.

[vii] Es gibt ein berühmtes Beispiel aus den USA, in dem eine Klageschrift auf sechs andere Fälle verweist, die allesamt nicht existieren: https://arstechnica.com/tech-policy/2023/05/lawyer-cited-6-fake-cases-made-up-by-chatgpt-judge-calls-it-unprecedented/.

[viii] https://www.br.de/nachrichten/netzwelt/chatgpt-ki-besteht-bayerisches-abitur-mit-bravour,TfB3QBw

[ix] https://www.businessinsider.com/list-here-are-the-exams-chatgpt-has-passed-so-far-2023-1#the-sat-3; zur Kritik an Benchmarks s. etwa https://aisnakeoil.substack.com/p/gpt-4-and-professional-benchmarks

[x] https://medium.com/@soltrinox/the-i-q-of-gpt4-is-124-approx-2a29b7e5821e

[xi] Eine nicht zu technische Einführung in die Technik hinter GPT findet sich bei https://writings.stephenwolfram.com/2023/02/what-is-chatgpt-doing-and-why-does-it-work/.

[xii] Eine Einführung in das Wesen algorithmischer Programme und dem Unterschied zum Maschinenlernen findet sich bei https://www.bidt.digital/was-ist-software/.

[xiii] ChatGPT basiert auf einem Sprachmodell der Firma OpenAI. Es gibt zahlreiche andere Modelle, wie etwa bei https://github.com/Mooler0410/LLMsPracticalGuide/blob/main/imgs/qr_version.jpg dargestellt.

[xiv] https://www.midjourney.com/home/, https://github.com/features/copilot; https://www.bing.com

[xv] Eine Einführung in das Training von ChatGPT findet sich bei https://build.microsoft.com/en-US/sessions/db3f4859-cd30-4445-a0cd-553c3304f8e2.

[xvi] https://theconversation.com/what-are-hollywood-actors-and-writers-afraid-of-a-cinema-scholar-explains-how-ai-is-upending-the-movie-and-tv-business-210360

[xvii] https://www.faz.net/pro/d-economy/chatgpt-und-ki-die-maechtigen-neuen-assistenzsysteme-18587321.html

[xviii] https://www.faz.net/pro/d-economy/endlich-neue-pruefungen-dank-chatgpt-18760199.html

[xix] S. etwa http://www.rechtzweinull.de/chatgpt-co-urheberrecht-bei-werken-der-kuenstlichen-intelligenz-ki-2/

[xx] https://blogs.microsoft.com/on-the-issues/2023/09/07/copilot-copyright-commitment-ai-legal-concerns/

[xxi] Z.B. https://fast-ai-movies.de/

[xxii] Eine Übersicht findet sich bei https://arxiv.org/abs/2310.03533.

[xxiii] https://github.blog/2022-09-07-research-quantifying-github-copilots-impact-on-developer-productivity-and-happiness/

[xxiv] https://www.forbes.com/sites/emmawoollacott/2023/08/17/chatgpt-has-liberal-bias-say-researchers/