Centrum für Europäische Politik
FREIBURG | BERLIN

cepAdhoc

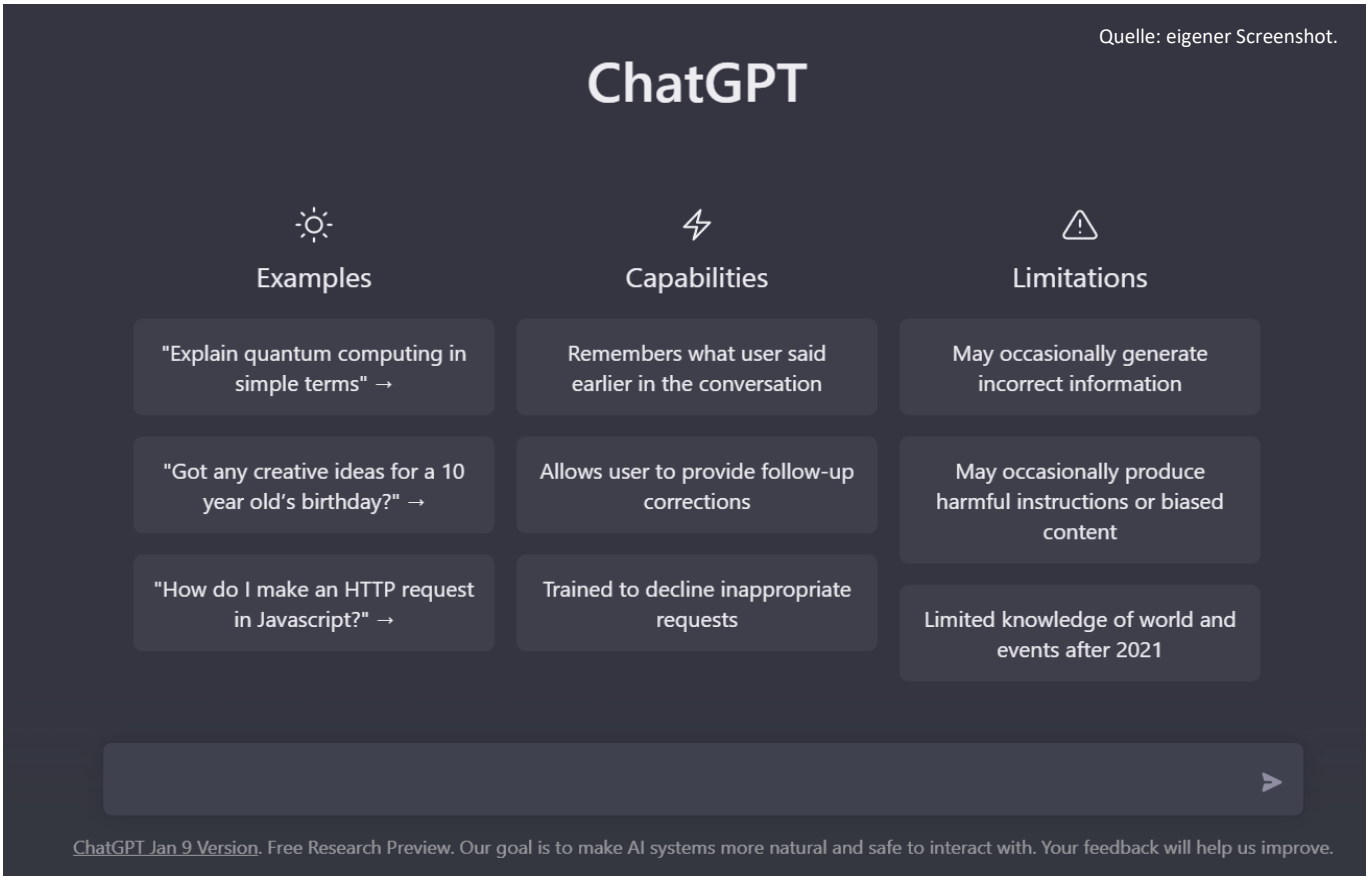
Nr. 1 | 2023 27. Januar 2023

ChatGPT erfordert mehr digitale Mündigkeit

Rahmenbedingungen für einen sicheren und produktiven Einsatz von generativen Sprachmodellen

Anselm Küsters

Quelle: eigener Screenshot.



Generative KI-Sprachsysteme wie ChatGPT sind in aller Munde. Die disruptive Technologie erfordert mehr gesellschaftsweite digitale Mündigkeit, denn nur mit einer informierten, reflektierten Nutzung können Verbraucher profitieren. Ohne sorgfältig gestaltete Rahmenbedingungen drohen hingegen politische Polarisierung und verschärfte Ungleichheit.

- Mit Hilfe vortrainierter und feinjustierter **Large Language Models** (LLMs) haben Forschende zuletzt den Stand der KI-Technik für viele Aufgaben revolutioniert, was **zahlreiche wohlstandsfördernde Anwendungsfälle** – zum Beispiel in den Bereichen Journalismus, Wissenschaft, Bildung und Informatik – ermöglicht.
- Ohne **Aufklärungskampagnen und regulatorisches Einwirken** droht allerdings eine Verschärfung negativer Entwicklungen im digitalen Raum, insbesondere **politische Polarisierung**, **ökonomische Monopolisierung** und **soziale Ungleichheit.**
- Aufgrund der Unsicherheit über den zukünftigen Innovationspfad sollten sich europäische Politik und geplante Gesetzesvorhaben wie das KI-Gesetz der EU auf die **Schaffung sensibler Rahmenbedingungen** beschränken. Dafür wird ein **5-Säulen-Modell** vorgeschlagen, das Transparenz der Modelle, Wettbewerb der Anbietenden, fairen Zugang, Standards zur Wahrung der Anwenderhoheit sowie den Schutz geistigen Eigentums und persönlicher Daten einfordert.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

1 Einleitung: der weltweite Siegeszug von ChatGPT

Seit dem 30. November 2022 fluten sie regelmäßig die Bildschirme von Nutzern Sozialer Medien – die Rede ist von Screenshots von Texten, die von dem Computerprogramm ChatGPT generiert wurden und sich lesen, als wären sie von einem Menschen geschrieben worden. Seit dieser Sprachbot damit begonnen hat, Sonette im Stile Shakespeares zu dichten, die Relativitätstheorie auf intuitive Weise zu erklären und neue Bierrezepte zu erfinden,[1] erregt das Potenzial generativer Künstlicher Intelligenz (KI) die Aufmerksamkeit der Öffentlichkeit. Nur wenige Tage nachdem ChatGPT von seinem Entwickler OpenAI, einem in San Francisco ansässigen Unternehmen, veröffentlicht wurde, probierten bereits mehr als eine Million Nutzer das Tool begeistert aus. Die aktuelle Revolution in generativen Sprachmodellen, so das Kernargument dieses cepAdhoc, erfordert eine neue Form von digitaler Mündigkeit, denn nur mit einer informierten und technisch reflektierten Nutzung resultieren echte Wohlstandsgewinne. Auf einer solchen Basis mögen ChatGPT und Co sogar eine Chance sein, die Marktmacht bestehender Plattformen zu brechen und Echokammern zu öffnen. Ohne Investitionen in eine spezifische digitale Mündigkeit und sorgfältig gestaltete Rahmenbedingungen droht hingegen eine Verschärfung negativer Trends wie politische Polarisierung, ökonomische Monopolisierung und soziale Ungleichheit.

ChatGPT basiert auf einem großen Sprachmodell (*Large Language Model*, oder LLM), das Mechanismen des maschinellen Lernens und eine Vielzahl an Trainingsdaten nutzt, um Nutzeranfragen mit menschlich wirkenden Texten zu beantworten. Es ist besonders effektiv bei der Generierung kohärenter, realistischer und intelligent klingender Texte in spezifischen Bereichen, die von vorprogrammierten Datenbanken oder Chatbots in der Regel nicht erfasst werden. Der Erfolg von OpenAI und seinem ChatGPT-Produkt hat LLMs beziehungsweise generative KI zu dem dominierenden Thema in der Technologiebranche für 2023 gemacht. Gleich zu Beginn des Jahres wurde OpenAI mit 29 Milliarden Dollar bewertet;[2] am 23. Januar verkündete Microsoft, dass es sich mit einem mehrjährigen „Multimilliarden-Dollar"-Investment beteiligen wolle.[3] Zum Vergleich: Im Jahr 2014 kaufte Google das KI-Unternehmen DeepMind, das OpenAI wohl am nächsten kommt, für lediglich 500 Millionen Dollar. Schon heute hat Microsoft begonnen, OpenAI-Technologie in seine Cloud-Computing-Dienste zu integrieren.[4] Die Unternehmensberatung McKinsey zählt angewandte KI – eine Rubrik, in die auch ChatGPT fällt – zu den momentan wichtigsten disruptiven Trends, deren Effekte gesellschaftsweite Relevanz besitzen.[5]

Während Anwender tagtäglich neue Experimente mit ChatGPT durchführen und Berater über das ökonomische Potenzial der Technologie sinnieren, sind Nutzen und Probleme der Technik noch immer umstritten, auch unter Experten. Die automatisch generierten Texte sind dermaßen klar strukturiert, recherchiert und referenziert, dass Bildung und Ausbildung wohl neu gedacht werden müssen.[6] Auch wenn ChatGPT nicht wirklich „versteht", was es schreibt, ist nicht klar, ob es als intelligenter Agent zu zählen ist und was Intelligenz von KI-Anwendungen überhaupt bedeutet. Die meisten Medienberichte konzentrieren sich bislang auf den Erfolg, ohne das Ausmaß und die Bandbreite der noch auftretenden

---

1 Manzullo, B. (2023), Atwater Brewery unveils new craft beer made using AI (freep.com) (18.01.2023).
2 Jin, B. und Kruppa, M.(2023), ChatGPT Creator Is Talking to Investors About Selling Shares at $29 Billion Valuation - WSJ (05.01.2023).
3 Knobloch, A. (2023). Microsoft investiert weitere Milliarden in ChatGPT-Entwickler OpenAI | heise online (21.01.2023).
4 Goldman, S. (2023), Microsoft Azure OpenAI service now generally available, with ChatGPT on the way | VentureBeat (January 16, 2023).
5 Chui, M. und Roberts, R. (2023), New year, new tech, no problem | McKinsey (17.01.2023).
6 Stokel-Walker, C. (2022), AI bot ChatGPT writes smart essays — should professors worry? (nature.com) (09.12.2022); Diebold, G. (2023), Higher Education Will Have to Adapt to Generative AI—And That's a Good Thing – Center for Data Innovation (17.01.2023).

Fehler ernsthaft zu diskutieren.[7] Während die Gesellschaft noch damit beschäftigt ist, die Auswirkungen von ChatGPT zu reflektieren, deuten Gerüchte darauf hin, dass das Nachfolgemodell GPT-4, das offensichtlich noch um Größenordnungen komplexer als ChatGPT sein wird, bereits fertig entwickelt ist und schon im Frühjahr 2023 erscheinen könnte.[8]

Das vorliegende cepAdhoc möchte in diesem Kontext Orientierung bieten und erste Leitlinien formulieren, die einen sicheren und produktiven Umgang mit der neuen Technologie gewährleisten, ohne zukünftige Innovationspfade im Bereich der generativen Sprachmodelle zu beschränken. Im Kern steht dabei die Frage, wie LLMs zu nutzen und zu regulieren sind, um ihr ökonomisches Potenzial zu entfalten und gleichzeitig ihre potenziellen Gefahren zu minimieren. Um dies zu beleuchten, wird in Abschnitt 2 zunächst die Philosophie von OpenAI und der technische Hintergrund großer Sprachmodelle vorgestellt. Auf dieser Basis wird dann das ökonomische Potenzial der Technik skizziert, das von der medizinischen Ausbildung bis hin zum Programmieren von Software-Code reicht (Abschnitt 3). Die probabilistische Natur von LLMs sorgt allerdings zugleich für zahlreiche Probleme, die – unbehandelt – drohen, negative Trends wie politische Polarisierung, ökonomische Monopolisierung und soziale Ungleichheit zu verstärken (Abschnitt 4). Daher plädiert das cepAdhoc für die Schaffung sensibler Rahmenbedingungen in Form eines 5-Säulen-Modells, das Transparenz der Modelle, Wettbewerb der Anbietenden, fairen Zugang, Mindeststandards sowie den Schutz geistigen Eigentums und persönlicher Daten einfordert (Abschnitt 5). Zuletzt werden die wichtigsten Ergebnisse zusammengefasst (Abschnitt 6).

2 Unternehmerischer und technischer Hintergrund

OpenAI wurde im Jahr 2015 von dem Tesla- und SpaceX-Unternehmer Elon Musk, dem Risikokapitalgeber Sam Altman und weiteren Partnern und Investoren als eine gemeinnützige Forschungsorganisation gegründet.[9] Basierend auf philosophischen Sichtweisen wie dem Effektiven Altruismus (*Effective Altruism*) und dem Langfristigkeitsdenken (*Longtermism*) beabsichtigten die Gründer mit OpenAI, intelligente KI-Systeme ohne finanziellen Gewinn auf eine Weise voranzutreiben, die der gesamten Menschheit zugutekommen kommen würde. Gesucht war nichts weniger als der heilige Gral der Computerwissenschaften, die sogenannte „künstliche allgemeine Intelligenz“ (*Artificial General Intelligence*, oder AGI), obwohl viele Experten glauben, dass diese vorerst unerreichbar bleiben und insbesondere nicht von großen Sprachmodellen herbeigeführt werden wird.[10] Im März 2019 wurde OpenAI in ein gewinnorientiertes Unternehmen verwandelt (OpenAI LP). Obwohl das Unternehmen immer noch vom Vorstand der ursprünglichen gemeinnützigen Organisation kontrolliert wird, können Investoren nun das bis zu Hundertfache ihrer Investitionen als Gewinn zurückerhalten. Auf OpenAI gehen bereits zahlreiche wichtige Produkte von generativer KI zurück, wie beispielsweise das Computerprogramm DALL-E, dass Bilder aus Textbeschreibungen generiert.[11] Doch erst mit ChatGPT erhielt das Unternehmen eine weltweite Aufmerksamkeit über den Kreis von Technologieexperten hinaus.

ChatGPT stellt die aktuelle Stufe in einem größeren Entwicklungsschub bei der Verarbeitung natürlicher Sprache dar, dem sogenannten *Natural Language Processing* (NLP). NLP beschreibt den Einsatz

---

7 Zum Beispiel: Roose, K. (2022), The brilliance and weirdness of ChatGPT | eKathimerini.com (07.12.2022).

8 Romero, A. (2021), GPT-4 Rumors From Silicon Valley (substack.com) (11.11.2022).

9 Für eine kurze Geschichte des Unternehmens, und insbesondere seine philosophischen Basis, siehe: Bernard, T. (2023), ChatGPT, Safety in Artificial Intelligence, and Elon Musk (techpolicy.press) (03.01.2023).

10 Für eine instruktive Diskussion zum aktuellen Stand von AGI, siehe: Marcus, G. und Booch, G. (2023), AGI will not happen in your lifetime. Or will it? (substack.com) (22.01.2023).

11 Für DALL-E, siehe: Slack, G. (2023), What DALL-E Reveals About Human Creativity (stanford.edu) (17.01.2023).

von Computertechniken zur Verarbeitung und Analyse großer Mengen von Daten in natürlicher Sprache, um die Interaktion zwischen Computern und Menschen zu verbessern, unter anderem durch Spracherkennung, Textzusammenfassung und die Identifikation bestimmter Entitäten.[12] Auf der Grundlage dieser NLP-Techniken können Sprachmodelle (*Language Models*, oder LMs) programmiert werden, die das nächste Wort in einem Satz auf der Grundlage der vorherigen Wörter vorhersagen. Die jüngsten Erfolge bei generativen KI-Anwendungen sind, neben einigen algorithmischen Verbesserungen, vor allem auf die zunehmende Größe dieser Sprachmodelle zurückzuführen, was gewöhnlicherweise an der Anzahl der Parameter und der Größe der Trainingsdaten („Tokens“) gemessen wird (Tabelle 1). Die Größe von State-of-the-Art-Sprachmodellen wächst aktuell jedes Jahr um den Faktor 10 oder mehr.[13] Darum spricht man auch von *Large Language Models* (LLMs).

Tab. 1: Zusammenfassung aktueller großer Sprachmodelle

Modell	Unternehmen	Parameter (in Mrd.)	Tokens (in Mrd.)	Ankündigung
GPT-4	*OpenAI*	*TBA*	*TBA*	Feb. 2023?
ERNIE-Code	Baidu	0.56	NA	Dez. 2022
ChatGPT	OpenAI	175	300	Nov. 2022
Galactica	Meta AI	120	450	Nov. 2022
BLOOMZ	BigScience	176	366	Nov. 2022
BLOOM	BigScience	176	350	Jul. 2022
Minerva	Google Research	540	818.5	Jun. 2022
PaLM	Google Research	540	780	Apr. 2022
Chinchilla	DeepMind	70	1400	Mär. 2022
Gopher	DeepMind	280	300	Dez. 2021
LaMDA	Google AI	137	168	Jun.2021
GPT-3	OpenAI	175	300	Mai 2020
Megatron-11B	Meta AI	11	2200	Apr. 2020
BERT	Google	0.3	137	Okt. 2018

Quelle: Daten von https://lifearchitect.ai/models/; eigene Auswahl. Bemerkung: „Tokens“ sind in diesem Kontext als die Anzahl an Wörtern im Trainingskorpus zu verstehen.

Es gibt verschiedene Arten von Sprachmodellen. Bei der neuesten Generation handelt es sich um sogenannte kontextualisierte Sprachmodelle, die Deep Learning verwenden, um für jedes Wort eine kontextspezifische Einbettung zu schätzen, d. h. eine Einbettung, die nicht nur das Wort erfasst, sondern auch die Beziehung, die es zu anderen Wörtern in der Umgebung hat.[14] Dies basiert auf der sprachtheoretischen Annahme, welche wohl bis Ludwig Wittgenstein zurückgeht, dass sich die semantische Bedeutung eines Wortes oftmals aus dem Kontext, in dem es verwendet wird, ablesen lässt. Der Trainingsprozess für ein auf Deep Learning basierendes Sprachmodell besteht darin, die Gewichte der Verbindungen zwischen den „Neuronen“ so anzupassen, dass das Modell die Wahrscheinlichkeit einer Wortfolge genau vorhersagen kann. Während ältere KI-Modellen im Sprachbereich darauf abzielten, Muster aus Text-Daten zu replizieren, geht es bei diesen neuen LLMs also darum, mithilfe von

[12] Für eine aktuelle, leicht zugängliche Übersicht zu NLP aus sozialwissenschaftlicher Sicht, siehe: Grimmer, J., Roberts, M.E. und Stewart, B.M. (2022), Text as Data. A New Framework for Machine Learning and the Social Sciences, Princeton: Princeton University Press.

[13] Li, C. (2020), OpenAI's GPT-3 Language Model: A Technical Overview (lambdalabs.com) (03.06.2020).

[14] Grimmer, J., Roberts, M.E. und Stewart, B.M. (2022), Text as Data. A New Framework for Machine Learning and the Social Sciences, Princeton: Princeton University Press, S. 88.

Wahrscheinlichkeitsrechnung ganz neuartige Wortfolgen zu erzeugen. Bekannte vortrainierte Modelle sind das von Open AI entwickelte GPT (*Generative Pre-trained Transformer*) und Googles BERT.

Mit Hilfe dieser vortrainierten Sprachmodelle und der Methodik der Feinabstimmung für bestimmte Aufgaben[15] haben Forschende in kürzester Zeit den Stand der Technik bei einer Vielzahl von Aufgaben erweitert, was anhand von bestimmten Benchmarks gemessen werden kann.[16] Hier ist allerdings zu berücksichtigen, dass es momentan noch an einheitlichen, holistischen Bewertungsstandards mangelt, was es Forschenden erschwert, die aktuelle Landschaft an Sprachmodellen klar und genau zu verstehen.[17] Die Wirksamkeit der neusten Generation von Sprachmodellen ist für KI-Experten dennoch in dreierlei Hinsicht überraschend.[18] Erstens skaliert die Leistung von LLMs eindeutig mit der Größe der Trainingsdatenmenge, was einen klaren Weg zu weiteren Verbesserungen suggeriert. Zweitens gibt es qualitative, menschlich klar wahrnehmbare Sprünge in der Leistungsfähigkeit, wenn die Modelle skalieren, was letztere nach Jahren der Forschung für verschiedene abgeleitete Anwendungskontexte (siehe Abschnitt 3) praktikabel macht. Drittens wird offensichtlich, dass viele Aufgaben, die dem Menschen Intelligenz abverlangen, mit einem ausreichend leistungsfähigen LLM auf die Vorhersage des nächsten Wortes reduziert werden können, sodass hier Effizienzvorteile entstehen. ChatGPT basiert auf der letzten öffentlichen Iteration von GPT (genauer gesagt auf einem Modell der GPT-3.5-Serie, deren Training Anfang 2022 abgeschlossen wurde) und ist daraufhin abgestimmt, Konversation besonders gut zu verstehen und auf eine natürliche und kontextgerechte Weise auf diese zu reagieren.

Das technische Ziel eines Sprachmodells ist folglich nicht, die beste oder wahrhafteste Antwort zu geben, sondern grammatikalisch und semantisch korrekten Sätzen hohe Wahrscheinlichkeiten zuzuordnen. Die KI-Forscherin Murray Shanahan vom Imperial College London betont, dass die Grundfunktion eines großen Sprachmodells außerordentlich vielseitig sei, aber dass jede Anwendung trotz dieser Vielseitigkeit letztlich auf einem Modell basiere, das nur eine einzige Aufgabe erfüllen könne, nämlich die Erzeugung statistisch wahrscheinlicher Fortsetzungen von Wortfolgen.[19] Diese probabilistische Natur von LLMs ermöglicht, dass Sprachanwendungen wie ChatGPT nicht einfach existierende Textdaten kopieren, sondern völlig Neuartiges erschaffen können. Letzteres wiederum führt zu zahlreichen kreativen Anwendungsmöglichkeiten für Verbraucherapplikationen mit großen potentiellen Wohlfahrtsgewinnen (Abschnitt 3), aber auch zu gravierenden Fehlern und Fehleinschätzungen der Technologie (Abschnitt 4).

3 Ökonomisches Potenzial

Generative Sprachmodelle erlauben verbesserte Konsumentenapplikationen, verkürzen die Entwicklungszeit derselben und machen leistungsstarke Funktionen für technisch nicht versierte Nutzer weit zugänglich.[20] So glaubt Erik Brynjolfsson, Direktor des Stanford Digital Economy Lab, dass ChatGPT „eine Menge routinemäßiger Arbeit beseitigen wird und gleichzeitig Menschen, die es verwenden,

[15] Die Feinabstimmung ist ein Prozess, bei dem ein Modell mit einigen neuen Daten für eine begrenzte Anzahl von Iterationen weiter trainiert und damit verbessert wird. Allerdings muss aufgepasst werden, ein *Over-fitting* zu vermeiden.

[16] Siehe: Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S., Schuh, P., Shi, K., Tsvyashchenko, S., Maynez, J., Rao, A., Barnes, P., Tay, Y., Shazeer, N., Prabhakaran, V. et al. (2022), PaLM: Scaling Language Modeling with Pathways, arXiv. https://doi.org/10.48550/arXiv.2204.02311.

[17] Dies ist die Kritik von: Liang, P., Bommasani, R., Lee, T., Tsipras, D., Soylu, D., Yasunaga, M., Zhang, Y., Narayanan, D., Wu, Y., Kumar, A., et al. (2022), Holistic evaluation of language models. arXiv preprint arXiv:2211.09110.

[18] Shanahan, M. (2022). Talking About Large Language Models. 10.48550/arXiv.2212.03551, S. 1.

[19] Shanahan, M. (2022). Talking About Large Language Models. 10.48550/arXiv.2212.03551, S. 4.

[20] Chui, M., Roberts, R. und Yee, L. (2022), How generative AI could change your business | McKinsey (20.12.2022).

möglicherweise in der Lage sein werden, kreativere Arbeit zu leisten."[21] Die Tatsache, dass heutige LLMs es erlauben, kreative Aufgaben mit KI anzugehen, ist die wohl wichtigste Neuerung dieser Technologiegeneration, die vor allem auf die Menge an Trainingsdaten und die enorme Erhöhung der Rechenleistung zurückzuführen ist.[22] Wie würden ChatGPT-betriebene Produkte und Dienstleistungen aussehen? Laut McKinsey gibt es Anwendungsmöglichkeiten im Marketing und Vertrieb (Erstellung von personalisierten Inhalten), Betrieb (Erstellung von Aufgabenlisten für die effiziente Ausführung einer bestimmten Tätigkeit), IT/Engineering (Schreiben, Dokumentieren und Überprüfen von Code), Risiko- und Rechtsabteilung (Beantwortung komplexer Fragen, Abrufen von großen Mengen rechtlicher Unterlagen und Verfassen und Überprüfen von Jahresberichten) sowie F&E (z.B. Beschleunigung der Arzneimittelentwicklung durch Entdeckung chemischer Strukturen).[23] Die Risikokapitalinvestitionen in generative KI beliefen sich im Jahr 2022 auf über 2 Milliarden Dollar.[24]

Die wohl meistdiskutierte Option bislang ist, dass ChatGPT als Front-End für Suchmaschinen eingesetzt werden könnte, um menschliche Anfragen zu beantworten. Eine KI-gestützte Suche würde bedeuten, dass man auf eine Frage nicht nur eine Liste von Links, sondern einen vollständigen, flüssigen Text mit der richtigen Antwort erhält. Unternehmen wie You.com und Neeva testen bereits Systeme, die die herkömmliche Internet-Suche mit großen Sprachmodellen à la ChatGPT verbinden.[25] Sundar Pichai, der aktuelle Google-CEO, hat daher die früheren Chefs und Gründer Larry Page und Sergey Brin zurückgeholt, um die KI-Strategie des Unternehmens im Kontext von ChatGPT zu aktualisieren.[26] Das Potenzial von generativen Sprachmodellen, Menschen dabei zu helfen, schneller und kohärenter zu schreiben, könnte auch Microsoft-Produkte enorm verbessern, was die oben erwähnte Beteiligung dieses Unternehmens an OpenAI erklärt. So könnten LLMs beispielsweise in Microsoft Word integriert werden, um es den Nutzern zu erleichtern, Dokumente zusammenzufassen oder neue Konzepte zu entwickeln. Auf ähnliche Weise könnten sie E-Mail-Programme mit besseren automatischen Vervollständigungsoptionen ausrüsten. Laut Experten werden große Sprachmodelle zudem bald in der Lage sein, auf Sprachbefehle zu reagieren oder Text laut vorzulesen, was insbesondere Menschen mit Lernschwierigkeiten oder Sehbehinderungen helfen würde.[27]

Neben diesen Anwendungsfällen im Software-Bereich, die sich recht zügig implementieren ließen, zeigen jüngste akademische Arbeiten, dass das Einsatzgebiet von generativen Sprachmodellen weit über reine Konsumentenapplikationen hinausgeht und diese auch entfernte, aber gesellschaftlich hochrelevante Gebiete wie die medizinische Ausbildung, Rechtsberatung oder Informatik bereichern können. So haben Forschende ChatGPT am Maßstab des *United States Medical Licensing Exam* gemessen, das aus drei Prüfungen besteht und von Medizinern in den USA abgelegt werden muss, um ärztlich tätig werden zu dürfen. ChatGPT erreichte bei allen drei Prüfungen ohne spezielles Training oder Fine-Tuning die Grenze, die zum Bestehen erforderlich war, und schien die gegebenen Antworten sogar verstehen zu können.[28] Diese Ergebnisse deuten darauf hin, dass große Sprachmodelle das Potenzial

[21] Aldrick, P. (2023), ChatGPT Will Be the Calculator for Writing, Top Economist Says - Bloomberg (18.01.2023), eigene Übersetzung.
[22] Chatterjee, M. (2023), 5 Questions for Zulfikar Ramzan - POLITICO (13.01.2023).
[23] Chui, M., Roberts, R. und Yee, L. (2022), How generative AI could change your business | McKinsey (20.12.2022).
[24] Clark, P.A. (2023), Generative artificial intelligence is driving tech's latest hype wave (axios.com) (10.01.2023).
[25] Voß, O. (2023), ChatGPT: Paradigmenwechsel für die Internetsuche - Tagesspiegel Background (19.01.2023).
[26] Weck, A. (2023), Furcht vor ChatGPT: Google holt Larry Page und Sergey Brin zurück (t3n.de) (21.01.2023).
[27] Heikkilä, M. (2023), Here's how Microsoft could use ChatGPT | MIT Technology Review (17.01.2023).
[28] Kung, T.H, Cheatham, M., Medenilla, A., Sillos, C., De Leon, L., Elepaño, C., Madriaga, M., Aggabao, R., Diaz-Candido, G., Maningo, J. und Tseng, V. (2022), Performance of ChatGPT on USMLE: Potential for AI-Assisted Medical Education Using Large Language Models, medRxiv 2022.12.19.22283643.

haben, bei der medizinischen Ausbildung und möglicherweise auch bei der klinischen Entscheidungsfindung zu unterstützen. Eine ähnliche Untersuchung zweiter Juraprofessoren sagt voraus, dass ein großes Sprachmodell bald in der Lage sein wird, den Multiple-Choice-Teil des *Multistate Bar Exam* zu bestehen, d.h. einen Bestandteil der juristischen Zulassungsprüfung in den USA, die ein angehender Anwalt dort ablegen muss.[29] GitHub Copilot, ein LLM-basiertes Vervollständigungstool für das Schreiben von Code, deutet bereits an, wie erfolgreiche Produkte in diesem Bereich auf der Grundlage generativer Sprachmodelle gebaut werden könnten. Experten loben dieses Produkt für seinen starken Produkt-/Markt-Fit, den immensen Mehrwert und die hohe Kosteneffizienz.[30] Einen Eindruck von der Potenz von GitHub Copilot vermittelt ein Werkstattbericht von Andrej Karpathy, dem früheren Direktor für KI bei Tesla, der zugibt, dass Copilot sein Coden „dramatisch beschleunigt" habe. Es sei nur schwer vorstellbar, nach diesem Technologiesprung zurück zu „manuellem Coden" zu gehen. Er schließt: „Ich lerne noch, es [GitHub Copilot] zu benutzen, aber es schreibt bereits ~80% meines Codes mit ~80% Genauigkeit. Ich kodiere nicht einmal wirklich, ich gebe nur Anweisungen und bearbeite."[31]

Trotzdem ist zu betonen, dass es noch eine gewisse Zeit dauern wird, bis diese Technologie den Weg von vielversprechenden Prototypen hin zu robusten Applikationen gemacht haben wird. Dies zeigen zum Beispiel erste Erfahrungen mit dem Einsatz von KI-Sprachmodellen zur Unterstützung der Kundenbetreuung. Aktuelle Untersuchungen zeigen, dass es in der Praxis der digitalen Kundenkommunikation noch immer zu gravierenden Problemen für Verbraucher kommt.[32] ChatGPT und Co bieten vor diesem Hintergrund zahlreiche Möglichkeiten, um Kundendienstmitarbeiter zu unterstützten, sei es in Form von verbesserten Chatbots, durch Audio- und Sprachverarbeitungen oder durch Hilfen inmitten eines Gesprächs, indem einem Mitarbeitenden passgenaue Informationen sekundenschnell zugespielt werden. Allerdings zeigt sich bei näherer Betrachtung, dass das Verstehen des Kunden, der Art seines Anliegens und insbesondere seiner emotionalen Verfasstheit bislang nur von Menschen überzeugend geleistet werden kann.[33] Letztere bleiben somit vorerst für einen guten und anpassungsfähigen Kundenservice unerlässlich. ChatGPT wirkt also eher als sinnvolles Komplement, denn als Substitut.

4 Problemkreise

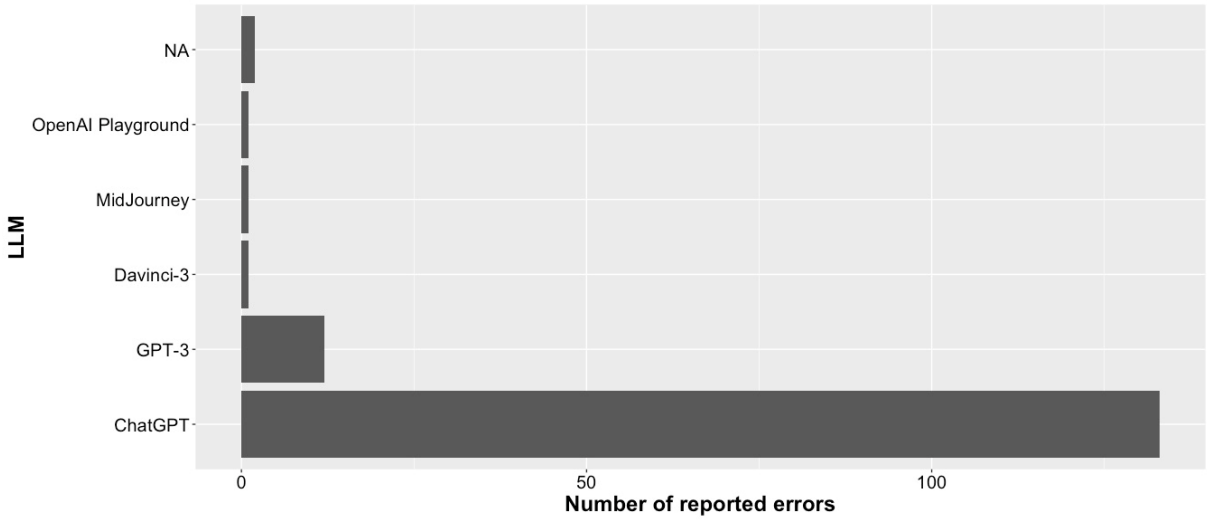
LLMs haben ein enormes Potenzial, Wissen zu demokratisieren und ökonomischen Wohlstand zu schaffen. Mit der Verbesserung generativer Sprachmodelle eröffnen sich neue Möglichkeiten in so unterschiedlichen Bereichen wie Gesundheitswesen, Recht und Informatik. Doch wie bei jeder neuen Technologie sollte man auch hier bedenken, wie sie missbraucht werden kann. Algorithmische Zusammenfassungen, die von ChatGPT und anderen LLMs produziert werden, können Fehler oder veraltete Informationen enthalten oder Nuancen und Unsicherheiten überdecken, ohne dass die Nutzer dies bemerken. Kritiker bemängeln, dass ChatGPT noch immer die gleichen Fehler wie seine Vorgängermodelle macht.[34] So kann ChatGPT zum Beispiel nicht zuverlässig zählen, die Reihenfolge der Ereignisse in einer Geschichte herausfinden, über die physikalische Welt nachdenken oder menschliche Denkprozesse mit ihrem Charakter in Verbindung bringen. Noch problematischer ist, dass Ergebnisse komplett erfunden sein können oder sexistische und rassistische Vorurteile aufweisen. Die fehlerhafte Logik von

[29] Bommarito, M.J. und Katz, D.M. (2022), GPT Takes the Bar Exam (29.12.2022), http://dx.doi.org/10.2139/ssrn.4314839.
[30] Dickson, B. (2022), GitHub Copilot is the first real product based on large language models (thenextweb.com) (10.07.2022).
[31] Andrej Karpathy auf Twitter (30.12.2022).
[32] Verbraucherzentrale Bundesverband (2023), Enttäschte Verbraucher-Erwartungen (vzbv.de) (18.01.2023).
[33] Chui, M. und Roberts, R. (2023), New year, new tech, no problem | McKinsey (17.01.2023).
[34] Heaven, W.D. (2022), ChatGPT is OpenAI's latest fix for GPT-3. It's slick but still spews nonsense. | MIT Technology Review (30.11.2022).

ChatGPT folgt direkt aus der oben beschriebenen Technik von LLMs, die ein plausibel klingendes Wort nach dem anderen produzieren, anstatt sich mit der Bedeutung der Sprache auf einer fundamentalen Ebene auseinanderzusetzen.

Es ist wichtig zu betonen, dass diese Probleme nicht nur bei ChatGPT auftreten, sondern bei generativen KI-Anwendungen im Allgemeinen. Riley Goodside, ein bekannter Experte für große Sprachmodelle, hat vor kurzem Claude, eines der neuesten LLMs, auf bekannte Fehlertypen hin untersucht. Seine detaillierte Analyse zeigt, dass das System immer noch schnell sogenannte „Halluzinationen" (also Falschaussagen) generiert, die insbesondere im Fall von ChatGPT von Experten stark kritisiert wurden.[35] Das Ausmaß und die Bandbreite an LLM-typischen Fehlern kann durch Auswertung einer Datenbank illustriert werden, die Ernest Davis, Professor für Computerwissenschaften an der New York University, jüngst mit Kollegen initiiert hat. Der Aufbau dieser Datenbank war verknüpft mit dem Aufruf an die KI-Community, sämtliche beim Experimentieren mit ChatGPT und anderen Sprachmodellen identifizierten Fehler einzutragen und, wenn möglich, zu klassifizieren. Zum Zeitpunkt der Erstellung dieses cepAdhoc (Mitte Januar 2023), also nur wenige Wochen nach Start des Vorhabens, umfasste diese Datensammlung bereits 150 gemeldete Fehler. Wie Abbildung 1 entnommen werden kann, beziehen sich die meisten gemeldeten Fehler auf ChatGPT (absolute Zahlen). Eine qualitative Durchsicht aller manuell eingegebenen Fehlerbeschreibungen zeigt, dass die meisten LLM-Irrtümer anwendungsübergreifend als Fehlinformation bzw. sachlicher Irrtum, fehlerhafte Logik, fehlerhafte Mathematik, fehlerhafte physikalische Argumentation oder als eine Mischung aus diesen Kategorien eingestuft wurden.

Abb. 1: Häufige ChatGPT/LLM-Fehler



Quelle: Eigene Darstellung, basierend auf Daten von: ChatGPT/LLM Errors (Public) - Google Tabellen.

Diese Fehler von LLMs sind nicht nur irritierend, sondern können potenziell negative gesellschaftliche Auswirkungen haben. Insbesondere besteht die Gefahr, dass LLMs Umweltschäden, ökonomische Ungleichheit und soziale Fragmentierung verschärfen könnten.[36] Das liegt, erstens, daran, dass die Verarbeitung von LLMs in Datenzentren erfolgt. Das Trainieren und Entwickeln dieser Modelle ist daher kostspielig, sowohl finanziell aufgrund der Kosten für Hardware, Strom und Cloud-Rechenzeit als auch ökologisch aufgrund des CO_2-Fußabdrucks, der erforderlich ist, um moderne Verarbeitungshardware

[35] Goodside, R. und Papay, S. (2023), Meet Claude: Anthropic's Rival to ChatGPT | Blog | Scale AI (17.01.2023).

[36] Klasky, E., Middha, A., Kim, S., Rosenfeld, H., Kleinman, M. und Parthasarathy, S. (2022), What's in the Chatterbox? Large Language Models, Why They Matter, and What We Should Do About Them, Research Report (April 2022), Ford School of Public Policy, https://stpp.fordschool.umich.edu/research/research-report/whats-in-the-chatterbox.

zu betreiben. Insbesondere kontextualisierte Worteinbettungen, wie sie im Falle von GPT oder BERT verwendet werden, sind kostspielig, erfordern effektiv den Einsatz von Grafikprozessoren (GPUs) und benötigen lange Zeit, unter Umständen sogar Monate, zum Trainieren.[37] So schätzen Forschende zum Beispiel, dass jede Trainingsinstanz von BERT 79 Stunden dauert und 4.000 – 12.000 Dollar an Cloud-Computing-Zeit kostet.[38] Angesichts der hohen Rechenkosten ist es schwierig, diese Sprachmodelle ohne großes Kapital zu reproduzieren.[39] Mögliche Auswege bieten kurzfristig zusätzliche Stufen des Vortrainings[40] oder Methoden der Inferenzoptimierung[41] und mittel- und langfristig die Quanteninformatik, die bereits mehrerer Ansätze des sogenannten *Quantum Natural Language Processing* (QNLP) erfolgreich getestet hat.[42] Moderne, auf Deep Learning basierende Ansätze der Sprachverarbeitung, wie das Trainieren von LLMs, könnten also durch quantenmechanische Ansätze ersetzt werden.

Hinzu kommt, zweitens, dass weder die zuständigen Entwicklungsteams noch die zugrundeliegenden Trainingsdaten marginalisierte Gemeinschaften in angemessener Weise repräsentieren. Das könnte dazu führen, dass LLMs die Meinungen und Erfahrungen dieser kleineren Gruppen systematisch minimieren oder falsch darstellen. In einem vielzitierten wissenschaftlichen Artikel warnen die bekannte KI-Ethikerin Timnit Gebru und ihre Mitarbeiter vor den Gefahren der Voreingenommenheit in großen Sprachmodellen und der Fehlinterpretation ihrer Ergebnisse.[43] Sie empfehlen, bei der Entwicklung von LLMs nicht nur die großen Umwelt- und Finanzkosten besser zu berücksichtigen, sondern auch mehr Ressourcen in die Kuratierung und sorgfältige Dokumentation von Datensätzen zu investieren sowie den Entwicklungsprozess großer Sprachmodelle an klare Forschungs- und Entwicklungsziele und Werte zu koppeln.

Nicht zuletzt kann die Verbreitung von LLM-basierten Suchmaschinen und Chatbots dazu führen, dass leichter glaubwürdige Desinformationen produziert werden können, die das Funktionieren demokratischer Systeme unterminieren. OpenAI-Forscher haben mit dem Center for Security and Emerging Technology der Georgetown University und dem Stanford Internet Observatory zusammengearbeitet, um zu untersuchen, wie große Sprachmodelle für Desinformationszwecke missbraucht werden könnten. Sie schlussfolgern, dass Sprachmodelle für Propagandisten in der Tat hoch nützlich sein werden und die Online-Manipulation wahrscheinlich verändern werden. Selbst wenn die fortschrittlichsten Modelle privat bleiben oder über eine Programmierschnittstelle kontrolliert werden, werden Propagandisten Open-Source-Alternativen nutzen können.[44] Ohne eine gewisse digitale Mündigkeit, die das Bedrohungspotenzial erkennt, kann es sogar sein, dass viele Menschen überhaupt nicht bemerken werden, wenn diese KI-Systeme verzerrte Inhalte oder Fehlinformationen generieren. Experimente

---

37 Grimmer, J., Roberts, M.E. und Stewart, B.M. (2022), Text as Data. A New Framework for Machine Learning and the Social Sciences, Princeton: Princeton University Press, S. 88.

38 Strubell, E., Ganesh, A. und McCallum, A. (2019), Energy and Policy Considerations for Deep Learning in NLP (Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics), S. 3645–3650.

39 Das wird insbesondere kritisiert von: Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M., Li, X., Lin, X. V., Mihaylov, T., Ott, M., Shleifer, S., Shuster, K., Simig, D., Koura, P. S., Sridhar, A., Wang, T. und Zettlemoyer, L. (2022). OPT: Open Pre-trained Transformer Language Models. arXiv. https://doi.org/10.48550/arXiv.2205.01068.

40 Wei, J. und Tay, Y. (2022), Better Language Models Without Massive Compute – Google AI Blog (googleblog.com) (29.11.2022).

41 Weng, L. (2023), Large Transformer Model Inference Optimization | Lil'Log (lilianweng.github.io) (10.01.2023).

42 Guarasci R., De Pietro G., Esposito M. (2022), Quantum Natural Language Processing: Challenges and Opportunities. In: Applied Sciences 12(11), S. 5651. https://doi.org/10.3390/app12115651.

43 Bender, E.M., Gebru, T., McMillan-Major, A. und Shmitchell S. (2021), On the dangers of stochastic parrots: Can language models be too big? (ACM Conference on Fairness, Accountability, and Transparency).

44 Goldstein, J.A., Sastry, G., Musser, M., DiResta, R., Gentzel, M. und Sedova, K. (2023), Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations, https://doi.org/10.48550/arxiv.2301.04246.

belegen, dass ChatGPT gefälschte Zusammenfassungen für wissenschaftliche Aufsätze schreiben kann, die Forscher nur schwer von denen unterscheiden können, die von Menschen geschrieben wurden.[45] Das erhöht die Wahrscheinlichkeit, dass diese falschen Ergebnisse weiterverbreitet werden. Verleger versuchen bereits, eine KI-Mitautorenschaft zu verbieten, da diese Systeme keine Verantwortung für den Inhalt und die Integrität ihrer Arbeit übernehmen können.[46] Insgesamt besteht die Gefahr, dass eine Simplifizierung oder Falschdarstellung wissenschaftlicher Ergebnisse durch LLMs informationelle Filterblasen verstärkt und damit politische Polarisierung begünstigt. Letztere gibt es zugegebenermaßen schon heute – aber die Möglichkeit, mit ChatGPT falsche Informationen in selbstbewusster und kontextsensitiver Sprache zu produzieren, potenziert die Möglichkeiten von Manipulatoren, kostengünstig und in hoher Geschwindigkeit plattformübergreifende Desinformationskampagnen zu führen.

Neben diesen breiten gesellschaftlichen Auswirkungen von LLMs bietet die Technologie natürlich auch Anwendungsmöglichkeiten in der Kriminalität, die ebenfalls von der Generierung überzeugender gefälschter Texte profitieren kann. In diesem Sinne wird ChatGPT die moderne Cyber-Bedrohungslandschaft revolutionieren, da es weniger qualifizierten Bedrohungsakteuren dabei hilft, Code für Cyber-Angriffe zu schreiben. In diesem Zusammenhang ist auch auf den oben erwähnten GitHub Copilot zu verweisen (Abschnitt 3). Experten von Check Point Research beschreiben, wie ChatGPT erfolgreich einen kompletten Infektionsablauf durchführen kann, von der Erstellung einer überzeugenden Spear-Phishing-E-Mail bis hin zur Ausführung einer Reverse Shell, die es Angreifern ermöglicht, einen Zielcomputer vollständig zu übernehmen.[47] Eine Untersuchung mehrerer Untergrund-Hacking-Communities durch die gleichen Experten zeigt zudem, dass es bereits erste Fälle gibt, in denen Cyberkriminelle OpenAI zur Entwicklung bösartiger Tools nutzten.

5 Ein 5-Säulen-Modell zur Förderung der digitalen Mündigkeit

Grundsätzlich ist bei Forderungen nach Regulierung generativer Sprachmodelle zu betonen, dass die Ergebnisse von ChatGPT bislang nur begrenzt bewertet werden können, da die wissenschaftliche Forschung zu LLMs noch weiter entwickelt werden muss und die proprietäre Natur von GPT externe Analysen erschwert.[48] Die Regulierung von KI-getriebenen Chatbots ist zudem komplex, weil sie im Prinzip zwei unterschiedliche Problemkreise verbinden muss, für die es jeweils unterschiedliche Lösungsansätze gibt.[49] Zum einen ist da die Verbreitung schädlicher KI-generierter Inhalte wie Hassreden, was automatisierte oder menschliche Methoden der *content moderation* nahelegt.[50] Zum anderen besteht die Gefahr von unbeabsichtigter Diskriminierung oder andere Verzerrungen, wenn Unternehmen generative Sprachmodelle für Einstellungsentscheidungen oder ähnliche Prozesse einsetzen, ohne die zugrundeliegenden Trainingsdaten und algorithmischen Logiken validiert und verstanden zu haben.

Anstelle eines rigorosen Verbots der neuen Technologie oder einer Detailregulierung, die zukünftige Innovationen in diesem sich rasant entwickelnden Sektor bremsen könnte, ist daher zunächst und

---

[45] Gao, C.A. et al. (2022), Comparing scientific abstracts generated by ChatGPT to original abstracts using an artificial intelligence output detector, plagiarism detector, and blinded human reviewers, Preprint bei bioRxiv https://doi.org/10.1101/2022.12.23.521610.

[46] Stokel-Walker, C. (2023), ChatGPT listed as author on research papers (nature.com) (18.01.2023).

[47] Check Point Research (2023), OPWNAI : Cybercriminals Starting to Use ChatGPT - Check Point Research (06.01.2023).

[48] Bommarito, M.J. und Katz, D.M. (2022), GPT Takes the Bar Exam (29.12.2022), http://dx.doi.org/10.2139/ssrn.4314839.

[49] Heikkilä, M. (2023), The EU wants to regulate your favorite AI tools (mailchi.mp) (09.01.2023).

[50] Wie Facebook gezeigt hat, ist es prinzipiell möglich – wenn auch nicht fehlerfrei – eine KI mit markierten Beispielen von Hassreden zu trainieren, um dann mit diesem Tool unmarkierte Formen der Toxizität automatisiert zu löschen. Diese Technik kann genutzt werden, um die Trainingsdaten von LLMs zu verbessern. Für das Säubern von ChatGPT-Trainingsdaten, siehe: Perrigo, B. (2023), OpenAI Used Kenyan Workers on Less Than $2 Per Hour: Exclusive | Time (18.01.2023).

grundlegend eine spezifische digitale Mündigkeit der Nutzenden notwendig. Gemeint ist damit eine informierte und technisch reflektierte Anwendung der neuen Sprachmodelle auf breiter gesellschaftlicher Ebene, da dies die entscheidende Voraussetzung für eine sichere und produktive Entwicklung der Technologie ist. Denkbar wären beispielsweise Aufklärungskampagnen sowie eine aktive Auseinandersetzung mit den Potenzialen und Problemen von Sprachmodellen in der Ausbildung von Schülern und Studenten. Murray Shanahan warnt zurecht, dass LLMs sich in ihrer Konstruktion zwar sehr vom Menschen unterscheiden, gleichzeitig aber so menschenähnlich in ihrem Verhalten sind, „dass wir genau darauf achten müssen, wie sie funktionieren, bevor wir über sie in einer Sprache sprechen, die menschliche Fähigkeiten und Verhaltensmuster suggeriert".[51]

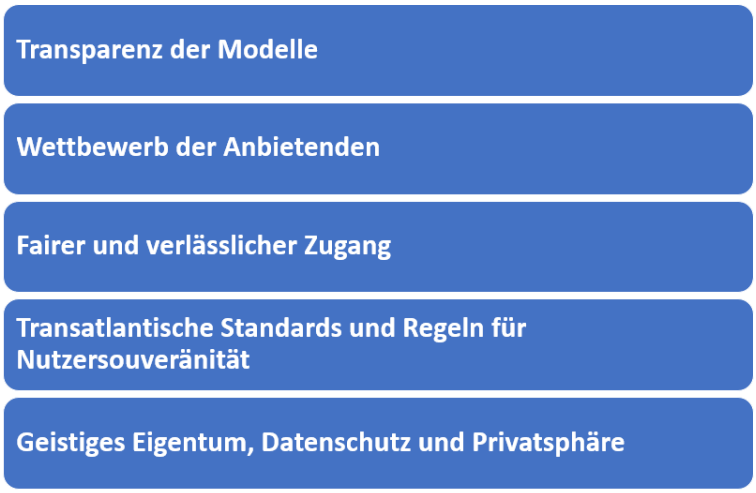
Neben Initiativen zur Erhöhung der digitalen Mündigkeit sollten sensible regulatorische und ökonomische Rahmenbedingungen geschaffen werden, um die angesprochenen Wohlfahrtsgewinne zu realisieren und die gesellschaftsweiten Gefahren zu minimieren. Denn Experten wie Jennifer King, Privacy and Data Policy Fellow in Stanford, gehen davon aus, dass viele Unternehmen trotz der zahlreichen mit ChatGPT verbundenen Risiken das Tool schnell integrieren werden.[52] Solche Rahmenbedingungen sind auf europäischer Ebene momentan noch nicht ausreichend aufgebaut: Der *Digital Markets Act* (DMA) zielt auf mächtige Plattform-Unternehmen, sogenannte „Gatekeeper", ab und definiert dafür hohe Schwellenwerte (6,5 Milliarden Euro Jahresumsatz, 45 Millionen aktive Nutzer pro Monat), die OpenAI noch lange Zeit nicht erfüllen wird (dies würde sich ändern, falls das Unternehmen von einem der BigTech-Unternehmen aufgekauft würde). Der *Digital Services Act* (DSA) enthält Verpflichtungen für Anbieter von digitalen Vermittlungsdiensten, worunter auch Suchmaschinen wie Google und Bing fallen. Dies würde relevant, falls Microsoft seine Pläne umsetzt, ChatGPT in Bing zu integrieren. Auch hier ist zu beachten, dass die Sorgfaltspflichten nach Art und Größe des Dienstes gestaffelt sind (die strengen DSA-Vorgaben gelten nur für sehr große Plattformen bzw. Suchmaschinen mit 45 Millionen aktiven Nutzern in der EU pro Monat). Am relevantesten erscheint daher das geplante KI-Gesetz (*AI Act*), welches die EU-Institutionen zurzeit aktualisieren möchten, um generative Sprachmodelle als sogenannte „Allzweck-KI"-Systeme zu erfassen, da diese für die verschiedensten Tätigkeiten eingesetzt werden können. Aufgrund der Komplexität und Undurchsichtigkeit dieser Systeme und der oben besprochenen Risiken ist diese Erweiterung der Regulierung zu begrüßen, um eine sichere Nutzung der neuen Technik zu gewährleisten.[53] Ein Ausschluss dieser Systeme aus dem KI-Gesetz würde hingegen die rechtliche Verantwortung auf die Nutzer der KI-Systeme statt auf die Entwickler verschieben, was insbesondere für europäische SMEs und Start-ups stark belastend wäre.[54] Allerdings sind konkrete Regeln für LLMs auf Basis des KI-Gesetzes nicht schnell zu erwarten: Der aktuelle Kompromiss sieht vor, die Kommission zu beauftragen, eine Folgenabschätzung und eine Konsultation durchzuführen, um die Vorschriften für „Allzweck-KI"-Systeme innerhalb von eineinhalb Jahren nach Inkrafttreten der Verordnung durch einen Durchführungsrechtsakt anzupassen.[55] Die vom Rat beschlossene Fassung des KI-Gesetzes sieht vor, dass alle generativen KI-Modelle, die für Hochrisikoanwendungen eingesetzt werden, alle Verpflichtungen des KI-Gesetzes für Hochrisikosysteme einhalten müssen – aber Juristen sind der Meinung, dass es unmöglich sein wird, ein umfassendes Risikomanagementsystem für alle möglichen Anwendungen solcher Systeme einzurichten.[56] Ein Änderungsantrag des Europäischen

[51] Shanahan, M. (2022). Talking About Large Language Models. 10.48550/arXiv.2212.03551, S. 3.
[52] Eadicicco, L. (2023), You'll Be Seeing ChatGPT's Influence Everywhere This Year - CNET (14.01.2023).
[53] So auch: Future of Life Institute (2022), General-Purpose-AI-and-the-AI-Act.pdf (artificialintelligenceact.eu) (Mai 2022).
[54] Siehe: ALLAI (2022), AIA-in-depth-Objective-Scope-and-Definition.pdf (allai.nl).
[55] Bertuzzi, L. (2022), EU countries adopt a common position on Artificial Intelligence rulebook – EURACTIV.com (06.12.2022).
[56] Hacker, P., Engel, A. und List, T. (2023), Understanding and Regulating ChatGPT – Verfassungsblog (20.01.2023).

Parlaments sieht zudem vor, dass Nutzer von LLMs offenlegen müssen, dass ihr Text von KI generiert wurde, es sei denn, Menschen überprüfen den Inhalt oder tragen die redaktionelle Verantwortung.[57]

Aufgrund der Unsicherheit über den zukünftigen Innovationspfad sollte sich die Politik auf fünf Säulen beschränken, die im Folgenden vorgestellt werden: Transparenz der Modelle, Wettbewerb der Anbietenden, Fairer Zugang, Hoheit über die inhaltlichen Grundlagen sowie Schutz des Geistigen Eigentums und privater Daten.[58] Abbildung 2 fasst dieses 5-Säulen-Modell für eine sichere und produktive Anwendung von generativen Sprachmodellen zusammen.

Abb. 2: 5-Säulen-Modell für eine sichere und produktive Anwendung von Sprachmodellen



Quelle: Eigene Darstellung.

1. **Transparenz der Modelle**: Obwohl es einen klaren Leistungsgewinn durch die ständige Erhöhung der Modellkapazität gibt, ist auch Experten nicht klar, was wirklich innerhalb der LLMs vor sich geht; diese bleiben eine *black box*.[59] Bei den wenigen LLMs, die über Programmierschnittstellen verfügbar sind, wird kein Zugang zu den vollständigen Modellgewichten gewährt, was ihre Untersuchung von außen erschwert.[60] Unklar bleibt auch, inwieweit Verzerrungen in den Trainingsdaten wirksam werden. Die fehlende Einsehbarkeit der bislang populären Modelle aus den USA erschwert es, diese für europäische Anwendungsfälle individuell anzupassen.[61] Für eine sichere und produktive Verwendung von LLMs ist daher zunächst Transparenz entscheidend. Entwickler von LLMs sollten darlegen, welche Trainingskorpora genutzt wurden und welche Logik von den verwendeten Algorithmen verfolgt wird.[62] Zudem zeigt das Beispiel ChatGPT, dass die neueste Generation an Sprachmodellen nicht nur von riesigen Mengen an Trainingsdaten abhängt, sondern von einer ebenso massiven Menge an menschlicher,

[57] Bertuzzi, L. und Stöckl, B. (2023), EU-Parlament will KI-Gesetz nachschärfen – EURACTIV.de (10.01.2023).

[58] Einige dieser Politikempfehlungen sind inspiriert von der Diskussion, die bei der Austauschveranstaltung „Chat GPT – Fluch oder Segen generativer KIs für die universitäre Lehre?“ am 17. Januar 2023 an der Humboldt-Universität zu Berlin stattfand, insbesondere die von Prof. Thorsten Hiltmann formulierten Rahmenbedingungen für eine positive Nutzung von ChatGPT wurden hier aufgegriffen. Siehe: Chat GPT – Fluch oder Segen generativer KIs? Digital History Berlin (hypotheses.org).

[59] Li, C. (2020), OpenAI's GPT-3 Language Model: A Technical Overview (lambdalabs.com) (03.06.2020).

[60] Das wird insbesondere kritisiert von: Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M., Li, X., Lin, X. V., Mihaylov, T., Ott, M., Shleifer, S., Shuster, K., Simig, D., Koura, P. S., Sridhar, A., Wang, T. und Zettlemoyer, L. (2022), OPT: Open Pre-trained Transformer Language Models, arXiv. https://doi.org/10.48550/arXiv.2205.01068.

[61] Saches, M. (2023), KI-Modelle: Wie die deutsche Antwort auf Chat GPT entstehen soll (faz.net) (24.01.2023).

[62] Klasky, E., Middha, A., Kim, S., Rosenfeld, H., Kleinman, M. und Parthasarathy, S. (2022), What's in the Chatterbox? Large Language Models, Why They Matter, and What We Should Do About Them, Research Report (April 2022), Ford School of Public Policy, https://stpp.fordschool.umich.edu/research/research-report/whats-in-the-chatterbox.

manueller Arbeit, die anscheinend notwendig ist, um unseriöse Informationen herausfiltern und die Anwendung so weniger toxisch zu machen.[63] Wie oben erwähnt sieht die EU aktuell vor, das geplante KI-Gesetz mit neuen Regeln zu aktualisieren, die generative Sprachmodelle als sogenannte „Allzweck-KI"-Systeme regulatorisch erfassen sollen. Laut Dragoș Tudorache, einem Mitglied der Liberalen im Europäischen Parlament, der das KI-Gesetz mitverhandelt, werden die Änderungen dahingehend wirken, dass Entwickler allgemeiner KI-Modelle zukünftig offener darüber sein müssen, wie ihre Modelle aufgebaut und trainiert werden.[64] Auf dieser Basis könnte die Forschungsgemeinschaft dann zu einem holistischeren Benchmarking aktuell populärer LLMs übergehen. Percy Liang, Professor für Informatik und Statistik, und Kollegen vom Center for Research on Foundation Models in Stanford haben dafür jüngst einen neuen Benchmarking-Ansatz, das sogenannte *Holistic Evaluation of Language Models* (HELM), vorgeschlagen, der darauf abzielt, die benötigte Transparenz zu schaffen.[65]

2. **Wettbewerb der Anbietenden**: Entscheidend ist, dass die Sprachmodelle von OpenAI keine dominierende Marktstellung erhalten, die ihre Programmierschnittstellen zu einem entscheidenden *bottleneck* für zahlreiche nachgeordnete Industrien machen. Ansonsten droht die Gefahr, dass sich ähnliche Monopolisierungsproblematiken wie bei den gegenwärtigen „Großen Fünf" (GAFAM) ergeben, denen nun mühsam nach vielen Jahren der legislativen Passivität mit Initiativen wie dem DMA und DSA begegnet werden muss. Auch der KI Bundesverband warnte jüngst vor der „Gefahr monopolartiger Strukturen" im Bereich der tiefen KI-Modelle.[66] In der sehr langen Frist ist eine solche „Winner-take-all"-Dynamik in generativer KI nicht zu erwarten, da alle Modelle auf ähnlichen Datensätzen mit ähnlichen Architekturen trainiert werden und sich Cloud-Anbieter technisch nicht stark voneinander unterscheiden, da sie dieselben Grafikprozessoren verwenden.[67] Wie in Tabelle 1 dargelegt, gibt es bereits mehrere erfolgreiche KI-Laboratorien, die die Forschung in diesem Bereich aktiv vorantreiben. Allerdings fallen ihre Erfolge bislang weitaus bescheidener aus als im Falle von ChatGPT. So musste die öffentliche Schnittstelle des von Meta entwickelten LLMs namens Galactica innerhalb kürzester Zeit zurückgezogen werden, da die akademische Öffentlichkeit feststellte, dass das Modell wissenschaftliche Fakten und Erkenntnisse falsch interpretiert hatte, bis hin zu Antisemitismus.[68] Für Experten ist Claude, das Sprachmodell von AnthropicAI, der momentan einzige echte Wettbewerber von ChatGPT,[69] was die Dringlichkeit für eine vielgestaltigere, kompetitive Angebotslandschaft unterstreicht. Zudem fällt auf, dass viele der aktuell prominenten Sprachmodelle auf Unternehmen wie Google, Facebook und Microsoft, die bereits als dominante BigTech-Unternehmen gelten, direkt oder indirekt zurückgehen (vgl. Tabelle 1). Sollten diese Modelle bzw. ihre Programmierschnittstellen zu einem wichtigen Input für nachgelagerte Industrien werden, wären daher wettbewerbspolitische Eingriffe notwendig – zu denken wäre zum Beispiel an eine Regulierung als „Gatekeeper" im Sinne des DMA.

---

[63] Im Falle von ChatGPT wurde ein großer Teil dieser Arbeit von schlecht bezahlten Arbeitskräften in Kenia geleistet. Siehe: Perrigo, B. (2023), OpenAI Used Kenyan Workers on Less Than $2 Per Hour: Exclusive | Time (18.01.2023).

[64] Heikkilä, M. (2023), The EU wants to regulate your favorite AI tools (mailchi.mp) (09.01.2023).

[65] Liang, P., Bommasani, R., Lee, T., Tsipras, D., Soylu, D., Yasunaga, M., Zhang, Y., Narayanan, D., Wu, Y., Kumar, A., et al. (2022), Holistic evaluation of language models, arXiv preprint arXiv:2211.09110.

[66] Saches, M. (2023), KI-Modelle: Wie die deutsche Antwort auf Chat GPT entstehen soll (faz.net) (24.01.2023).

[67] So die Analyse von: Bornstein, M., Appenzeller, G. und Casado, M. (2023), Who Owns the Generative AI Platform? | Andreessen Horowitz (a16z.com) (19.01.2023).

[68] Heaven, W.D. (2022), Why Meta's latest large language model only survived three days online | MIT Technology Review (18.11.2022).

[69] Goodside, R. und Papay, S. (2023), Meet Claude: Anthropic's Rival to ChatGPT | Blog | Scale AI (17.01.2023).

3. **Fairer und verlässlicher Zugang**: Um weltweite und inner-gesellschaftliche Ungleichheiten nicht zu verschärfen ist ein fairer und verlässlicher Zugang zu dieser neuen Wissensressource notwendig. Dies ist umso dringender, als OpenAI seit kurzem einen bezahlten Zugang zu ChatGPT vorsieht (voraussichtlich 42 Dollar pro Monat), der einen besseren Dienst verspricht.[70] Im aktuellen Kontext einer multipolaren, zunehmend fragmentierten Weltwirtschaft, die von geopolitischer Rivalität auch und gerade im Digitalbereich geprägt ist, kommt erschwerend hinzu, dass laut Schätzungen des KI Bundesverbandes 73 Prozent der mit Deep Learning trainierten LLMs aus den USA kommen, 15 Prozent aus China.[71] Um einen fairen Zugang zu garantieren und strategische Abhängigkeiten zu vermeiden, könnte beispielsweise die Entwicklung eines großen europäischen LLM mit öffentlichen Geldern unterstützt werden, dessen Trainingsdaten zudem Werte reflektieren, die der EU besonders wichtig sind und die die beschriebenen *biases* zu vermeiden suchen. Häufig von der EU genannte KI-Werte beziehen sich auf demokratische Grundsätze wie Fairness oder Verantwortlichkeit. Technisch wäre dies umsetzbar, indem man beispielsweise LLMs auf bestimmten öffentlich zugänglichen wissenschaftlichen Artikeln in einer großen Vielfalt von Bereichen trainiert.[72] In der Tat gibt es schon ein in San Francisco ansässigen Unternehmens, das LLMs in der wissenschaftlichen Forschung erprobt und dafür ein Tool zur Beantwortung von Fragen anhand der gängigen Literatur entwickelt. Für Deutschland hat der KI Bundesverband mit einer Machbarkeitsstudie bereits Ideen für eine Infrastruktur zur Schaffung von großen KI-Modellen entwickelt und die von ihm ins Leben gerufene Initiative „Large European AI Model“ (LEAM) drängt auf eine dedizierte Supercomputing-Infrastruktur in Europa, mit der LLMs in Zukunft heimisch und CO_2-neutral trainiert werden können.[73] Auch wenn solche Maßnahmen auf den ersten Blick radikal erscheinen, bedeutet Passivität letztlich, dass man sich langfristig von opaken Sprachmodellen abhängig macht, deren implizite Werturteile von außen nicht nachzuvollziehen sind. Dieses Argument basiert auf der Annahme, dass LLMs wie ChatGPT tatsächlich die revolutionäre Wirkung entfalten werden, die Forschende und Unternehmensberater momentan vorhersagen. Eine weniger kostenintensive Alternative könnte BLOOM sein, ein offen zugängliches Sprachmodell mit 176 Milliarden Parametern (vgl. Tabelle 1). Dieses wurde von einer größeren Forschungsgemeinschaft als Reaktion darauf entworfen, dass die meisten LLMs von ressourcenstarken Organisationen entwickelt und häufig der Öffentlichkeit vorenthalten werden.[74]
4. **Standards und Regeln zur Gewährleistung der Nutzersouveränität**: Bislang gibt es kaum systematische Regelungen oder Mechanismen zur Aufrechterhaltung von Standards für moderne KI-Anwendungen. Die EU hat mit dem erwähnten KI-Gesetz nun erstmals Rechtsvorschriften vorgeschlagen, während die US-amerikanische Regierung nur unverbindliche Prinzipien vorsieht. Zudem droht eine regulatorische Fragmentierung und damit Rechtsunsicherheit: Laut einer Analyse von Holistic AI sind sich Kanada, Großbritannien, die OECD, die EU und Kalifornien nicht einig, welche Systeme in den Anwendungsbereich der diversen KI-Regulierungen

[70] Leitinger, R. (2023), ChatGPT Professional: Preis für die Premium Version geleaked (robert-leitinger.com) (22.01.2023).
[71] Saches, M. (2023), KI-Modelle: Wie die deutsche Antwort auf Chat GPT entstehen soll (faz.net) (24.01.2023).
[72] Für eine ähnliche Forderung im Hinblick auf die US National Science Foundation, siehe: Klasky, E., Middha, A., Kim, S., Rosenfeld, H., Kleinman, M. und Parthasarathy, S. (2022), What's in the Chatterbox? Large Language Models, Why They Matter, and What We Should Do About Them, Research Report (April 2022), Ford School of Public Policy, https://stpp.fordschool.umich.edu/research/research-report/whats-in-the-chatterbox.
[73] Bienert, J. (2023), Der Weg zu großen KI-Modellen in Deutschland - Tagesspiegel Background (23.01.2023).
[74] Scao, T. L., Fan, A., Akiki, C., Pavlick, E., Ilić, S., Hesslow, D., Castagné, R., Luccioni, A. S., Yvon, F., Gallé, M., Tow, J., Rush, A. M., Biderman, S., Webson, A., Ammanamanchi, P. S., Wang, T., Sagot, B., Muennighoff, N. et al. (2022), BLOOM: A 176B-Parameter Open-Access Multilingual Language Model, arXiv. https://doi.org/10.48550/arXiv.2211.05100.

fallen sollten.[75] Immerhin hat das jüngste Treffen hochrangiger EU- und US-Beamter im Rahmen des Trade and Technology Council zur Verabschiedung einer „gemeinsamen Roadmap für vertrauenswürdige KI und Risikomanagement" geführt, die anstrebt, ein gemeinsames Regelwerk für die Entwicklung dieser neuen Technologie zu erstellen.[76] Ein Mangel an systematischen Regelungen und Standards wurde von Forschenden zuletzt speziell für LLMs beklagt.[77] Neben der oben angesprochenen Weitergabe von mehr Informationen über Sprachmodelle an Nutzende des Produktes sollten die Entwickler generativer KI dazu verpflichtet werden, bestimmte Beschränkungen für die von den Modellen erzeugten Ergebnisse einzubauen, um *Hate Speech* zu unterbinden, ihre Ergebnisse zu überwachen, um der Verbreitung von Desinformation entgegenzuwirken, und missbräuchliche Nutzer zu sperren. Trotz der signifikanten Verringerung schädlicher und falscher Ergebnisse, die die ChatGPT-Entwickler durch den Einsatz von menschlichem Feedback im Trainingsprozess erreichen konnten, räumt OpenAI ein, dass sein Modell immer noch toxische und verzerrte Ergebnisse erzeugen kann.[78] Im Sinne der angesprochenen LEAM-Initiative ist auch an europäische Standards für das Trainieren von LLMs zu denken, die hohe Ansprüche an Datenschutz, transparente Algorithmen und CO2-Neutralität bei diesem Prozess vorschreiben.[79] Nutzer von generativen KI-Sprachmodellen sollten wiederum verpflichtet werden, den Einsatz solcher Chatbots und die spezifischen Eingabeaufforderungen (*prompts*), die zu ihrem Resultat führten, zu dokumentieren, wenn sie die Inhalte weiterpublizieren, zum Beispiel in Form eines wissenschaftlichen Aufsatzes.[80]

5. **Geistiges Eigentum, Datenschutz und Privatsphäre**: Schließlich gibt es komplexe, bislang unbeantwortete Fragen bezüglich des Schutzes von geistigem Eigentum, der Privatsphäre und des Datenschutzes. Bezeichnenderweise hat Amazon seine Mitarbeiter davor gewarnt, Geheimnisse mit ChatGPT zu teilen, da nicht klar ist, wie das System vertrauliche Unternehmensdaten verwendet.[81] Die Datensätze, die generativen KI-Modellen zugrunde liegen, werden typischerweise aus dem Internet gescrapt, ohne dass die Zustimmung lebender Künstler oder urheberrechtlich geschützter Werke eingeholt wird.[82] Wenn diese Modelle dann aufgefordert werden, Kunstwerke im Stil eines noch lebenden Künstlers zu erstellen, ohne dass dafür eine Lizenz gewährt wurde, ist dies aus urheberrechtlicher Sicht problematisch. Sollten alle Autoren von Trainingsdaten ihre Zustimmung geben müssen oder eine Entschädigung erhalten? Diese Frage ist besonders brisant, da einige Beobachter davon ausgehen, dass gerade kreative Arbeitskräfte durch generative KI ersetzt werden.[83] Außerhalb der EU gibt es bereits Sammelklagen, bei denen Betroffene gegen einen Anbieter von KI-generierten Bildern vorgehen.[84]

Eine Umsetzung der hier gelisteten Rahmenbedingungen könnte das Vertrauen der Bevölkerung in die neue Technologie verstärken und damit die Innovation schneller umsetzbar machen kann. Zugleich werden durch ein besseres Verständnis von generativen Sprachmodellen und ihre regulatorische

---

75 Gulley, A. und Hilliard, A. (2023), Lost in Transl(A)t(I)on: Differing Definitions of AI (holisticai.com) (16.12.2022).
76 TTC Joint Roadmap for Trustworthy AI and Risk Management | Shaping Europe's digital future (europa.eu).
77 Klasky, E., Middha, A., Kim, S., Rosenfeld, H., Kleinman, M. und Parthasarathy, S. (2022), What's in the Chatterbox? Large Language Models, Why They Matter, and What We Should Do About Them, Research Report (April 2022), Ford School of Public Policy, https://stpp.fordschool.umich.edu/research/research-report/whats-in-the-chatterbox.
78 Aligning Language Models to Follow Instructions (openai.com).
79 Saches, M. (2023), KI-Modelle: Wie die deutsche Antwort auf Chat GPT entstehen soll (faz.net) (24.01.2023).
80 Stokel-Walker, C. (2023), ChatGPT listed as author on research papers (nature.com) (18.01.2023).
81 Kim, E. (2023), Amazon Warns Staff Not to Share Confidential Information (businessinsider.com) (24.01.2023).
82 Salkowitz, R. (2022), Midjourney Founder David Holz On The Impact Of AI On Art (forbes.com) (17.09.2022).
83 Salkowitz, R. (2022), AI Is Coming For Commercial Art Jobs. Can It Be Stopped? (forbes.com) (19.09.2022).
84 Meineck, S. (2023), Sammelklage: Streit um Bild-Generatoren soll vor Gericht landen (netzpolitik.org) (19.01.2023).

Einbettung diffuse Ängste über eine Ersetzung menschlicher Kreativität genommen, denn wie oben dargelegt deuten erste Erfahrungen mit der nach wie vor fehleranfälligen Technologie darauf hin, dass diese eher als ökonomisches Komplement denn als Substitut wirken wird.

6 Fazit

Mit ChatGPT, und insbesondere der für 2023 erwarteten vierten Iteration des zugrundliegenden Sprachmodells (GPT-4), ist die Zukunft der Konsumentenapplikationen eingetroffen. Schon jetzt lassen sich zahlreiche wohlfahrtssteigernde Anwendungsfälle absehen, unter anderem in der Informationsbeschaffung, dem Programmieren von Software-Code und der Medizin. Gleichwohl muss die hohe Anzahl und die Bandbreite von Fehlfunktionen, die ChatGPT auch gegenüber seinen Vorgängerversionen nicht korrigieren konnte, mehr in das öffentliche Bewusstsein gerückt werden. So besteht die Gefahr, dass LLMs Umweltschäden, ökonomische Ungleichheit und soziale Fragmentierung letztlich verschärfen. Durch die gesteigerte Glaubwürdigkeit von Desinformationskampagnen droht eine politische Polarisierung. Hinzu kommen neue Gefahren im Cyberschutz, da Kriminelle die Sprachbots zur schnellen und kostengünstigen Erstellung von schädlichen Angriffscodes verwenden können.

Dieses cepAdhoc hat argumentiert, dass die potenziellen Gefahren minimiert und die Wohlfahrtsgewinne für europäische Verbraucher erschlossen werden können, wenn die Entwicklung generativer KI von Anfang an und auf gesellschaftlich breiter Basis informiert und reflektiert begleitet wird. In anderen Worten, es braucht eine breitenwirksame digitale Mündigkeit, die durch technische Aufklärungskampagnen und Bildungsangebote zu fördern ist. Dabei helfen auch die vorgestellten Rahmenbedingungen, die die entstehende EU-Digitalregulatorik um DMA, DSA und AI Act sinnvoll ergänzen und das Potenzial haben, das Vertrauen der Bevölkerung in die neue Technologie zu verstärken, womit die technologische Umsetzung letztlich beschleunigt werden kann. Das hier vorgestellte 5-Säulen-Modell für eine sichere und produktive Anwendung von generativen Sprachmodellen fordert eine bessere Transparenz der Modelle, einen verstärkten Wettbewerb der Anbietenden, einen fairen und zuverlässigen Zugang zu dieser neuen Wissensressource, Anwenderhoheit über inhaltliche Grundlagen sowie den Schutz geistigen Eigentums und persönlicher Daten. Auf einer solchen Basis können LLMs sogar eine Chance sein, die Marktmacht bestehender Plattformen zu brechen und Echokammern zu öffnen.

Durch ein besseres Verständnis der Technik und seine regulatorische Einbettung können auch die momentan zirkulierenden, diffusen Ängste über eine Ersetzung menschlicher Arbeitskraft und Kreativität genommen werden, die die Weiterentwicklung generativer KI-Systeme langfristig behindern. Dass solche Ängste verfehlt sind, wurde schon Anfang des Jahres in einer symptomatischen Episode deutlich, als ein Anhänger des Songwriters Nick Cave ChatGPT dazu nutzte, um ein Lied im Stil seines Idols zu generieren. Nick Cave empfand das Endprodukt als zutiefst unbefriedigend und verwies dabei auf einen zentralen Faktor des kreativen Schaffensprozesses, der auch durch immer größere Sprachmodelle nicht verschwinden wird: “ChatGPT hat kein inneres Wesen, es ist nirgendwo gewesen, es hat nichts ertragen, es hat sich nicht getraut, über seine Grenzen hinauszugehen und hat daher auch nicht die Fähigkeit zu einer gemeinsamen transzendenten Erfahrung, da es keine Grenzen hat, über die es hinausgehen könnte.”[85] Das hätte auch ChatGPT nicht schöner schreiben können.

[85] Eigene Übersetzung aus dem Englischen: https://www.theredhandfiles.com/chat-gpt-what-do-you-think/.

Centrum für Europäische Politik
FREIBURG | BERLIN

Autor:

Dr. Anselm Küsters, LL.M., Leiter des Fachbereichs Digitalisierung und Neue Technologien
kuesters@cep.eu

Centrum für Europäische Politik FREIBURG | BERLIN
Kaiser-Joseph-Straße 266 | D-79098 Freiburg
Schiffbauerdamm 40 Raum 4315 | D-10117 Berlin
Tel. + 49 761 38693-0